Do we really need to catch them all? A new User-guided Social Media Crawling method

Fredrik Erlandsson[1,*], Martin Boldt[1,+], Piotr Bródka[2,+], and Henric Johnson[1]

[1]Blekinge Institute of Technology, Department of Computer Science and Engineering, Sweden
[2]Wrocław University of Science and Technology, Department of Computational Intelligence, Poland
[*]fredrik.erlandsson@bth.se
[+]these authors contributed equally to this work

ABSTRACT

With the growing use of popular social media services like Facebook and Twitter it is hard to collect all content from the networks without access to the core infrastructure or paying for it. Thus, if all content cannot be collected one must consider which data are of most importance. In this work we present a novel User-guided Social Media Crawling method (*USMC*) that is able to collect data from social media, utilizing the wisdom of the crowd to decide the order in which user generated content should be collected, to cover as many user interactions as possible. *USMC* is validated by crawling 160 Facebook public pages, containing 368 million users and 1.3 billion interactions, and it is compared with two other crawling methods. The results show that it is possible to cover approximately 75% of the interactions on a Facebook page by sampling just 20 % of its posts, and at the same time reduce the crawling time by 53%. What is more, the social network constructed from the 20% sample has more than 75% of the users and edges compared to the social network created from all posts, and has very similar degree distribution.

Supplementary Information

S1 Table. Detailed data used to construct Fig. 3, including the standard deviation for both Interactions and Crawling time. Please note that all data are normalized.

Type	Size	Interactions	(std)	Time	(std)
USMC - Number of likes	0.050	0.507	0.266	0.257	0.200
USMC - Number of comments	0.050	0.401	0.269	0.250	0.201
USMC - Post life time	0.050	0.254	0.279	0.149	0.166
Chronological	0.050	0.143	0.141	0.112	0.147
Random	0.050	0.049	0.022	0.049	0.034
USMC - Number of likes	0.100	0.621	0.247	0.350	0.219
USMC - Number of comments	0.100	0.510	0.268	0.339	0.224
USMC - Post life time	0.100	0.362	0.290	0.232	0.196
Chronological	0.100	0.240	0.176	0.202	0.211
Random	0.100	0.101	0.034	0.101	0.050
USMC - Number of likes	0.200	0.745	0.200	0.467	0.227
USMC - Number of comments	0.200	0.627	0.248	0.455	0.234
USMC - Post life time	0.200	0.524	0.273	0.361	0.202
Chronological	0.200	0.402	0.205	0.338	0.238
Random	0.200	0.200	0.046	0.200	0.068
USMC - Number of likes	0.400	0.867	0.125	0.647	0.202
USMC - Number of comments	0.400	0.785	0.167	0.630	0.194
USMC - Post life time	0.400	0.718	0.201	0.566	0.177
Chronological	0.400	0.624	0.190	0.546	0.231
Random	0.400	0.400	0.054	0.400	0.082
USMC - Number of likes	0.600	0.945	0.063	0.792	0.146
USMC - Number of comments	0.600	0.892	0.105	0.771	0.155
USMC - Post life time	0.600	0.860	0.131	0.741	0.141
Chronological	0.600	0.787	0.147	0.716	0.187
Random	0.600	0.600	0.054	0.599	0.082
USMC - Number of likes	0.800	0.982	0.026	0.906	0.080
USMC - Number of comments	0.800	0.950	0.062	0.886	0.098
USMC - Post life time	0.800	0.944	0.058	0.877	0.091
Chronological	0.800	0.905	0.102	0.868	0.119
Random	0.800	0.801	0.043	0.801	0.064
USMC - Number of likes	0.900	0.992	0.015	0.953	0.048
USMC - Number of comments	0.900	0.975	0.037	0.944	0.048
USMC - Post life time	0.900	0.973	0.033	0.940	0.053
Chronological	0.900	0.957	0.064	0.939	0.068
Random	0.900	0.899	0.035	0.899	0.055
USMC - Number of likes	0.950	0.996	0.011	0.975	0.034
USMC - Number of comments	0.950	0.987	0.020	0.972	0.027
USMC - Post life time	0.950	0.987	0.018	0.970	0.036
Chronological	0.950	0.977	0.053	0.970	0.042
Random	0.950	0.950	0.024	0.950	0.033

S2 Table. Facebook pages statistics. If page does not have a name it means that the name is not in Latin characters set or does not exist anymore.

Name	Page id	Posts	Users	Comments	Likes	Edges	Nodes
Royal Club Consulting	203841476411447	562	182	38	1,300	3	25
Sustainable Development Policy & Practice / P...	126385310775420	1,756	273	37	384	4	27
Ski-Akademie Schladming	266340220123141	254	949	64	1,773	12	44
Say NO to Bullying	246881265345291	372	469	129	791	34	46
Chaddsford Winery	493075535051	207	480	160	814	101	96
488878434469215	488878434469215	578	1,889	224	4,872	107	140
Posthotel Schladming	184073274591	156	600	162	1,806	138	74
24-Hour Bully Stake-Out	150499681635588	1,218	750	1,008	1,485	344	264
Dourakis Winery	106096217726	281	3,048	400	7,042	360	183
Energy Saver	121219973056	704	2,909	644	6,397	381	448
Play Station, Nitendo, X Boxs E.t.c	10237302714	605	390	143	385	403	120
Betty Boom	9564135890	493	657	648	3,365	658	167
Charlie's Sub Sandwich Station	268228483234	1,126	1,555	1,118	3,685	704	482
THE nightrace Schladming official site	220818858019	256	3,085	628	4,932	945	410
Keep Matt Free	192646367542354	140	3,634	402	5,435	1,767	279
Tiger	11132188727	1,163	1,579	499	2,161	2,251	325
WINERY	146839468660306	749	4,659	1,063	15,979	2,320	491
Cancer Research UK Relay For Life	6700961047	667	1,328	881	3,258	2,631	387
Sharrott Winery	15554410418	1,013	1,585	1,463	6,880	2,851	425
Drink Air	310752782271834	642	2,271	1,417	5,771	3,018	636
Mutt Lynch Winery	161976601328	614	3,170	1,227	7,018	3,242	626
NEWS24	282571518441363	2,228	26,868	3,459	138,183	3,975	1,710
Festiwal Polskich Filmów Fabularnych	251797246747	688	5,207	2,443	14,210	5,040	901
Stop It Now!	16871848717	1,152	1,286	1,276	3,565	5,113	389
ATOMIC SKIING	132728146765723	1,445	22,089	1,946	58,695	5,545	1,279
PLAY STATION 2	18651374805	2,900	1,877	1,302	1,671	5,619	913
Caring and Courageous	178435075560368	4,523	13,182	4,577	55,642	5,805	1,192
Rumah Madani - Busana Muslim	111452573760	9,391	18,291	18,571	19,623	9,494	3,500
Farmville Bonus	130833830320642	638	16,052	2,775	33,285	10,230	2,203
Occupy Wall Street	258106877558909	7,584	14,227	6,600	23,867	18,254	3,032
Friends of NRA	261098770903	1,742	68,063	4,588	112,618	18,960	3,301
Drew Curtis' Fark.com	42145905998	2,763	13,603	5,775	22,681	24,740	3,506
15755709838	15755709838	1,075	23,161	5,981	92,914	34,855	3,247
Chateau Elan Winery & Resort	19286773617	1,131	25,008	4,814	41,928	36,065	3,950
Bicky Burger	9727196242	623	1,908	2,579	1,920	37,888	636
Hold My Hand	157264947680409	8,275	45,548	35,242	226,703	57,844	7,378
Oceana	6334782252	2,153	78,971	6,969	180,096	63,346	4,488
Orange	7558451780	14,736	31,107	26,559	44,124	73,516	9,518
One Year Bible Blog	30775112500	4,409	50,652	16,744	339,679	76,305	5,729
Pureology Serious Colour Care	16792472126	3,326	15,936	8,428	31,866	82,857	4,924
Rainforest Action Network	8002590959	5,185	118,060	16,067	249,414	83,752	8,816
Intel Malaysia	144511635581827	3,408	123,276	12,407	207,804	102,349	6,868
Remove and Replace Every Teapublican in 2012	152358668171541	3,627	25,755	9,773	91,126	103,286	4,971
AMBER Alert	226402156107	3,160	90,175	8,751	143,833	108,954	6,075
Pakistan Air Force	169629069729559	2,516	75,821	18,155	201,082	150,473	11,126
Smilebox	6289201895	6,194	48,109	11,578	74,251	157,524	7,095
Intel Australia	145393495515570	1,144	46,891	11,222	63,045	159,249	8,063
MGM Studios	18962571991	7,461	13,356	10,548	60,500	161,193	3,252
A Little Inspiration	344267635599110	11,919	442,244	49,308	1,751,077	180,823	30,482
99%	295474380481252	1,390	49,500	30,334	134,141	181,617	8,887
5676133521	5676133521	35,602	41,559	15,718	46,018	205,499	9,536
Disneynature	41120927270	429	67,000	10,875	129,282	229,551	7,265
What Would You Do?	10689860898	538	13,222	11,274	36,033	233,914	5,575
kate spade new york	10737653985	1,187	104,558	18,568	278,173	235,796	11,750
I Relate Feelings	166785736790960	2,791	219,182	28,564	1,057,158	254,768	17,995
CNNMoney	6651543066	11,238	120,824	57,311	348,925	260,575	15,064
Mall of America	6898166798	9,144	87,679	27,967	146,402	261,074	13,887
HTCampus.com	174009145975879	3,344	28,727	10,669	52,768	264,058	6,727
Clarins	98029424570	6,118	36,072	18,895	81,851	292,926	9,416
Intel France	186901908014467	1,607	32,898	16,398	39,914	298,323	10,143
JAPAN4YOU	99489209116	5,966	117,584	42,984	382,144	303,821	17,141
Disaster Response on Facebook	250083749935	6,414	79,809	14,234	94,800	366,692	10,739
I Love My Family	148167278107	9,165	618,079	37,131	1,141,265	387,171	24,332
NDTV Goodtimes	76026867998	5,630	355,869	35,962	832,376	393,171	23,646
News24 Kenya	201728286564087	6,172	53,767	31,945	52,681	397,015	23,489
We The People	178463258907212	57,499	161,707	91,279	831,279	439,834	30,766
Office	178191330369	4,973	196,862	28,155	252,763	462,469	18,707
335574863819	335574863819	236	71,616	13,139	118,502	476,668	9,810

Continued on next page

S2 Table. (continued)

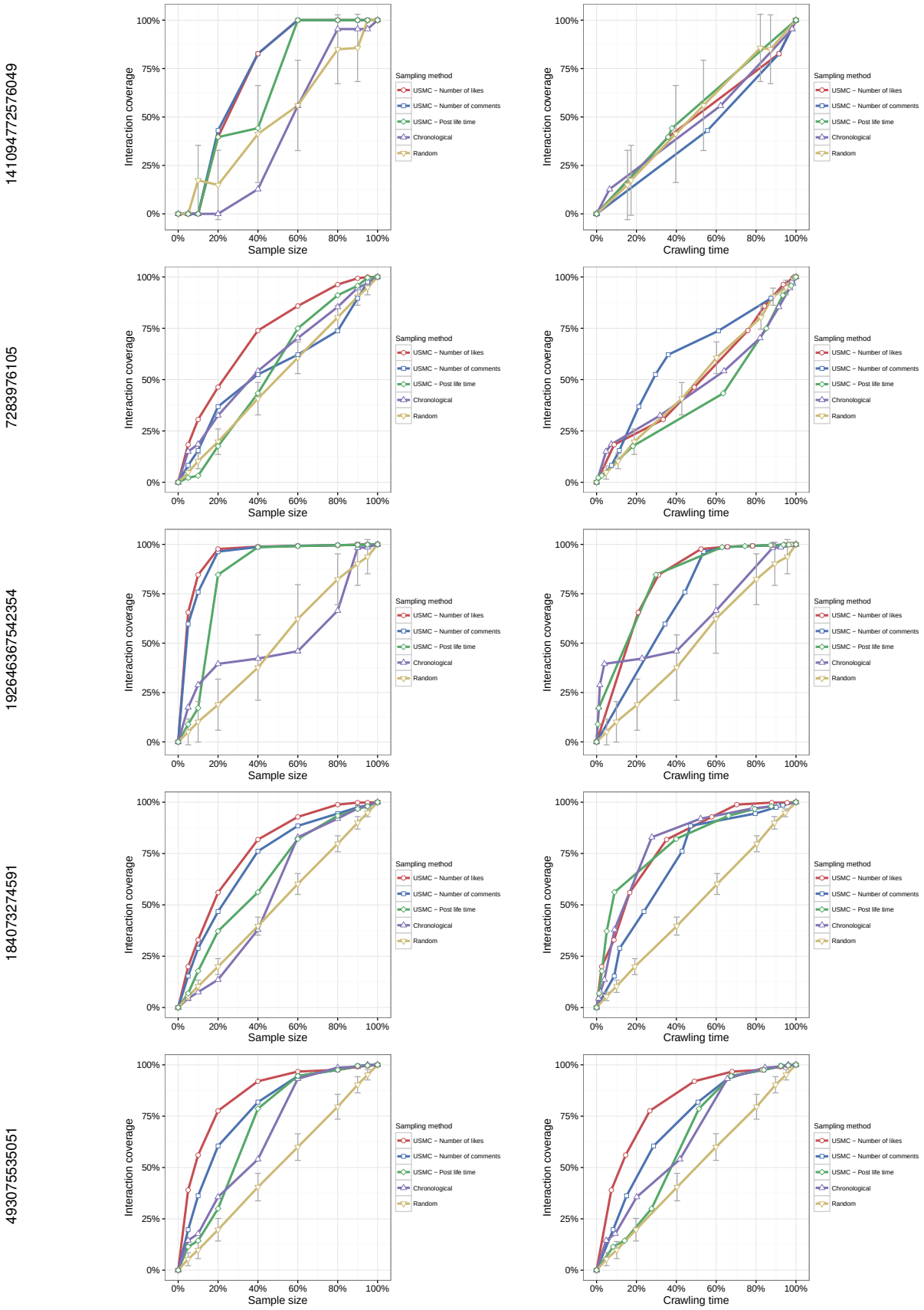
Name	Page id	Posts	Users	Comments	Likes	Edges	Nodes
Neutrogena	7717041259	7,878	98,269	33,064	161,540	518,038	18,512
Genevieve Gorder	174431941806	2,473	41,105	19,784	97,135	583,291	10,175
TurboTax	7511533723	13,322	46,976	65,687	80,095	608,739	14,050
(RED)	6829493713	9,014	928,159	44,807	1,368,432	666,520	37,841
Sierra Club	6204742571	3,804	134,502	21,287	488,584	703,952	12,101
Occupy Seattle	254620607914006	39,616	298,041	112,906	862,086	730,694	48,446
Nigeria	7283976105	66	17,218	10,698	17,385	863,555	7,846
Thich Nhat Hanh	7691064634	1,316	854,688	35,984	1,487,600	888,384	27,649
Liberal Democrats	5883973269	1,306	27,725	46,378	107,284	1,035,517	11,618
Human Target	83431257070	229	19,990	15,921	49,295	1,309,384	8,038
ONE	11055104471	10,758	553,321	55,342	1,474,206	1,330,108	38,716
Diet Coke	8605796091	13,778	117,444	24,389	181,455	1,332,970	16,178
Lifehacker	7568536355	11,727	247,459	139,009	1,049,963	1,638,928	51,873
Defenders of Wildlife	5720973755	9,660	111,035	59,268	740,886	1,742,348	21,188
First Bank of Nigeria Limited	218502344846561	1,537	100,883	61,248	168,091	2,006,432	38,655
Sony Music Brasil	99219862934	1,352	168,946	60,921	464,592	2,013,773	34,346
125787337496837	125787337496837	10,781	39,337	84,484	420,257	2,043,807	14,672
Madison Square Garden	28859306498	7,208	783,949	56,885	1,129,189	2,362,698	44,147
MetLife	164505093569983	4,646	294,975	54,785	1,208,807	2,388,519	29,902
WellDoneStuff.Com	302896416450257	3,546	1,325,457	117,839	3,085,706	2,427,758	96,841
Hilton HHonors	120963705931	5,843	228,520	28,670	373,519	2,500,219	19,915
The BULLY Project	107214895991663	13,934	393,667	64,502	858,783	2,707,881	37,157
Avaaz	8340223883	23,173	992,140	86,337	1,892,211	2,750,809	61,139
Amazon Books	150864087963	5,523	125,512	37,304	200,124	2,898,421	25,579
Sigur Rós	6187954123	16,735	224,333	42,050	585,551	2,942,991	27,348
NBC Politics	154957517930371	5,007	76,351	43,401	148,379	2,944,870	25,753
Liftopia	8114223318	3,973	47,001	50,065	158,699	3,198,393	18,983
Nat Geo Wild	196008660929	1,598	4,080,404	96,961	6,853,487	3,325,417	82,676
199299723534097	199299723534097	19,287	231,256	33,437	385,775	3,340,919	24,590
Human Rights Watch	42940254353	49,177	676,432	116,163	1,140,654	4,057,011	63,191
Royal Air Force	26035834884	14,825	518,911	124,547	2,117,085	4,194,305	55,807
141094772576049	141094772576049	7	14,708	10,123	12,877	4,502,259	5,746
JAPAN AIRLINES (JAL)	195152223850783	1,097	673,490	89,594	4,026,970	4,545,436	46,168
Amazon Student	135155166518704	5,286	181,508	43,497	200,777	4,647,731	38,497
San Francisco | The Official Guide	55139614218	13,305	735,183	116,336	2,054,348	4,806,264	80,978
California Right To Know	178713385549168	4,462	644,007	70,345	1,247,518	5,128,532	50,749
193072760827	193072760827	10,141	534,199	66,006	2,660,263	5,196,490	35,186
Sprint	8389383510	121,624	402,589	421,677	569,390	5,302,697	83,940
Joss Stone	5630135837	19,542	253,053	64,049	797,799	5,347,896	36,272
*Well, if that isn't the skank calling the who...	292440510802935	8,516	532,138	132,555	2,384,824	6,319,615	62,416
AirAsia	18801397386	103,808	318,869	331,450	790,496	7,622,130	92,077
Newsweek	18343191100	15,499	743,737	187,464	1,214,252	9,070,226	117,362
Epica	8031842923	1,487	364,197	110,159	1,813,942	10,855,119	57,131
Love	135922659784530	15,197	1,088,422	190,854	5,808,937	15,925,918	120,750
Ted	317928348241564	23,608	309,082	81,892	506,756	16,871,827	61,995
fluid ink	121128974566421	1,721	443,794	51,213	544,276	19,096,755	38,826
humor inteligente	292974040727244	7,252	6,368,200	520,868	15,957,356	19,231,726	418,409
KFC	7144906559	45,228	300,844	136,044	395,669	22,211,434	91,923
U.S. Soccer	32421823940	22,984	949,625	283,819	2,569,209	22,726,254	132,817
HBO	113408673932	32,476	2,895,356	226,018	3,525,002	23,898,618	188,232
ThinkGeek	6399067073	30,368	822,694	225,890	4,221,585	24,058,867	108,818
Alanis Morissette	6002796793	616	378,711	86,702	1,164,302	24,684,390	50,190
Planned Parenthood Action	8934429638	22,338	706,756	300,004	3,234,499	27,142,303	82,602
Royal Pains	141512340315	7,855	179,121	88,729	580,721	29,647,131	40,745
CBS	47360808996	48,503	5,904,875	406,287	7,607,876	31,024,271	323,247
No, Microsoft Word, I'm pretty sure I know how...	104375499654834	12,963	1,816,142	160,176	6,801,222	32,206,661	108,276
True Activist	129370207168068	17,664	2,851,690	289,156	6,947,506	43,402,257	207,935
CNBC	97212224368	15,440	1,432,933	440,391	2,306,032	53,416,686	255,087
AT&T	8576093908	231,521	1,330,859	957,877	2,251,109	54,860,445	238,830
Jennifer Hudson	7886660434	437	1,247,861	164,138	1,971,608	59,743,420	128,195
The Richard Dawkins Foundation for Reason and ...	8798180154	33,559	810,320	483,200	4,417,444	60,627,254	130,476
So You Think You Can Dance	9748634303	85,185	889,270	293,600	2,580,510	61,042,994	142,010
Boston Celtics	8725012666	2,148	831,423	378,473	3,737,903	62,757,635	170,571
The Office	6092929747	65,507	2,067,280	368,369	8,404,195	64,391,891	207,428
KHUSUS MUSLIMAH	108919785796003	6,726	659,541	157,370	1,484,280	71,203,208	117,501
Jeep	7037526514	88,571	332,389	374,401	822,363	71,275,779	137,159
Guns N' Roses	10901008068	1,398	922,449	374,576	2,358,218	85,765,225	220,459
Thirty Seconds to Mars	5618127822	3,298	1,768,814	411,789	12,625,270	87,803,406	185,361
USA TODAY	13652355666	20,932	7,451,142	1,097,625	13,976,321	94,570,053	694,618

Continued on next page

S2 Table. (continued)

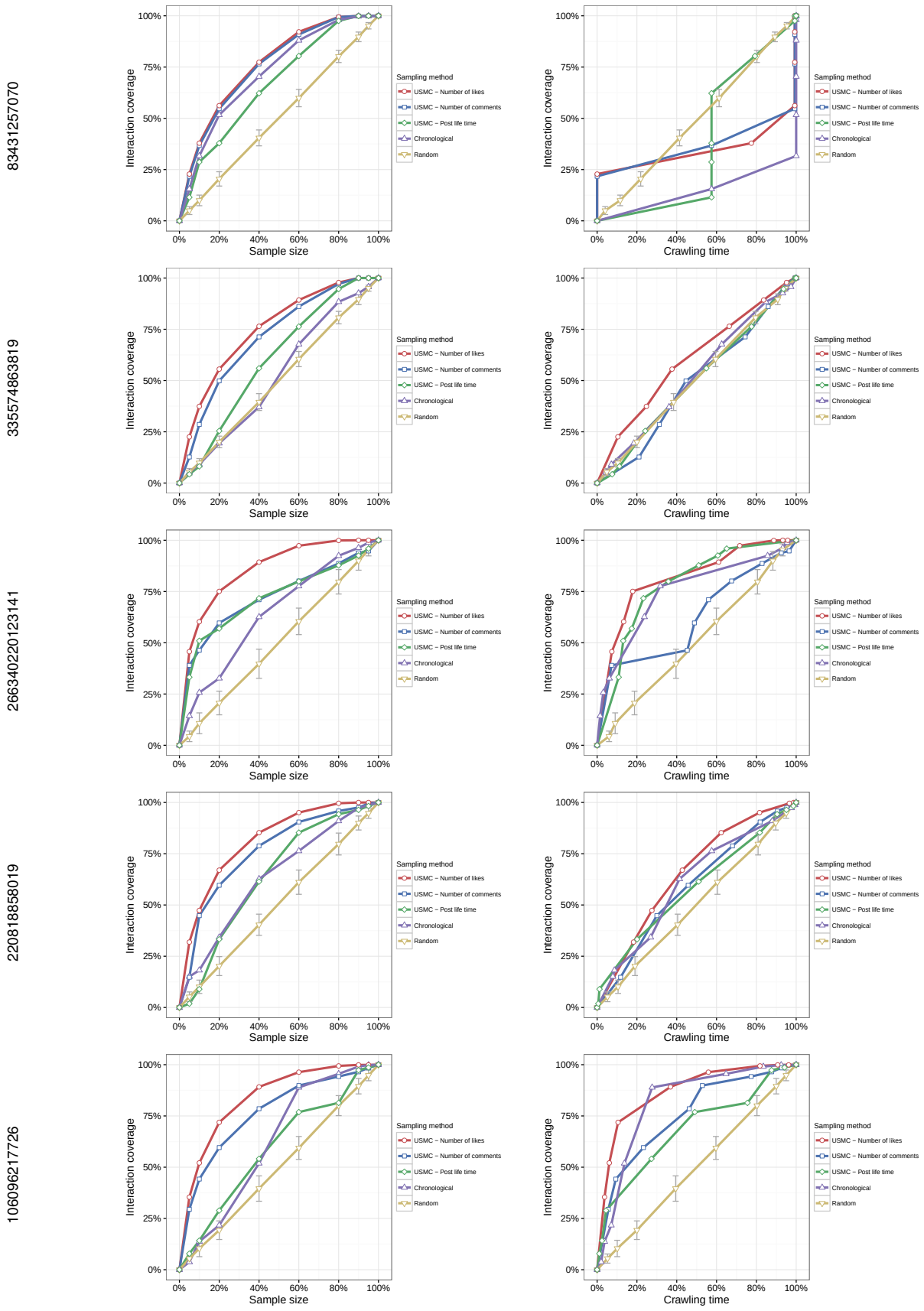
Name	Page id	Posts	Users	Comments	Likes	Edges	Nodes
The Wall Street Journal	8304333127	29,186	9,333,492	1,636,732	20,822,568	96,838,683	925,938
Kelly Rowland	5485793674	46,629	1,912,487	242,766	5,639,775	149,022,735	155,241
I Love Being Black	18493706927	5,670	3,660,378	687,000	16,818,408	152,389,438	358,607
Le Monde.fr	14892757589	33,787	11,698,482	2,486,741	34,930,046	180,750,060	1,181,501
Islamic Book	192169930808550	33,784	6,946,485	1,288,775	23,140,984	184,572,975	792,567
SID	17382788643	1,380	573,048	771,087	3,039,529	185,595,986	204,982
Katy Perry	7126051465	470,528	2,187,154	669,443	3,683,168	229,310,526	411,750
Do Something	7630216751	12,451	1,340,787	381,540	2,615,132	237,194,443	261,412
Mr. Bean	17774451468	533	3,017,698	553,378	4,118,601	246,789,355	367,959
NPR	10643211755	13,869	4,198,783	1,436,847	11,264,175	264,604,989	573,617
The Onion	20950654496	15,864	8,600,978	1,607,803	28,288,510	277,474,993	767,184
PoliticsNation with Al Sharpton	280920811923248	80,829	2,742,367	2,336,604	13,597,984	484,549,856	668,319
Buffalo Wild Wings	9418270899	78,156	2,357,805	902,121	6,147,686	617,537,397	514,143
The New York Times	5281959998	46,858	35,982,349	5,329,607	87,662,740	705,053,685	3,020,786
Being Conservative	134193140910	836	1,223,614	973,422	3,899,235	741,039,627	423,040
Linkin Park	8210451787	275,143	2,871,517	1,347,061	7,170,784	744,469,926	681,054
LA Lakers	144917055340	3,041	3,653,065	1,682,734	29,295,627	771,588,768	599,813
MLB	5768707450	9,439	2,946,677	2,602,902	25,145,970	1,088,696,317	533,729
Rush Limbaugh	136264019722601	3,870	9,554,690	2,398,526	31,336,180	1,240,262,427	1,414,824
CNN	5550296508	466,942	8,376,715	5,146,195	27,090,373	1,434,229,849	1,458,146
The White House	63811549237	6,742	32,581,271	9,238,522	78,841,712	4,238,052,189	2,385,752
The Tea Party†	133279166727577	55,016	77,084,290	27,550,352	308,495,988		
Barack Obama†	6815841748	4,147	113,379,978	13,250,519	275,134,360		

† Page too big to generate SNA.



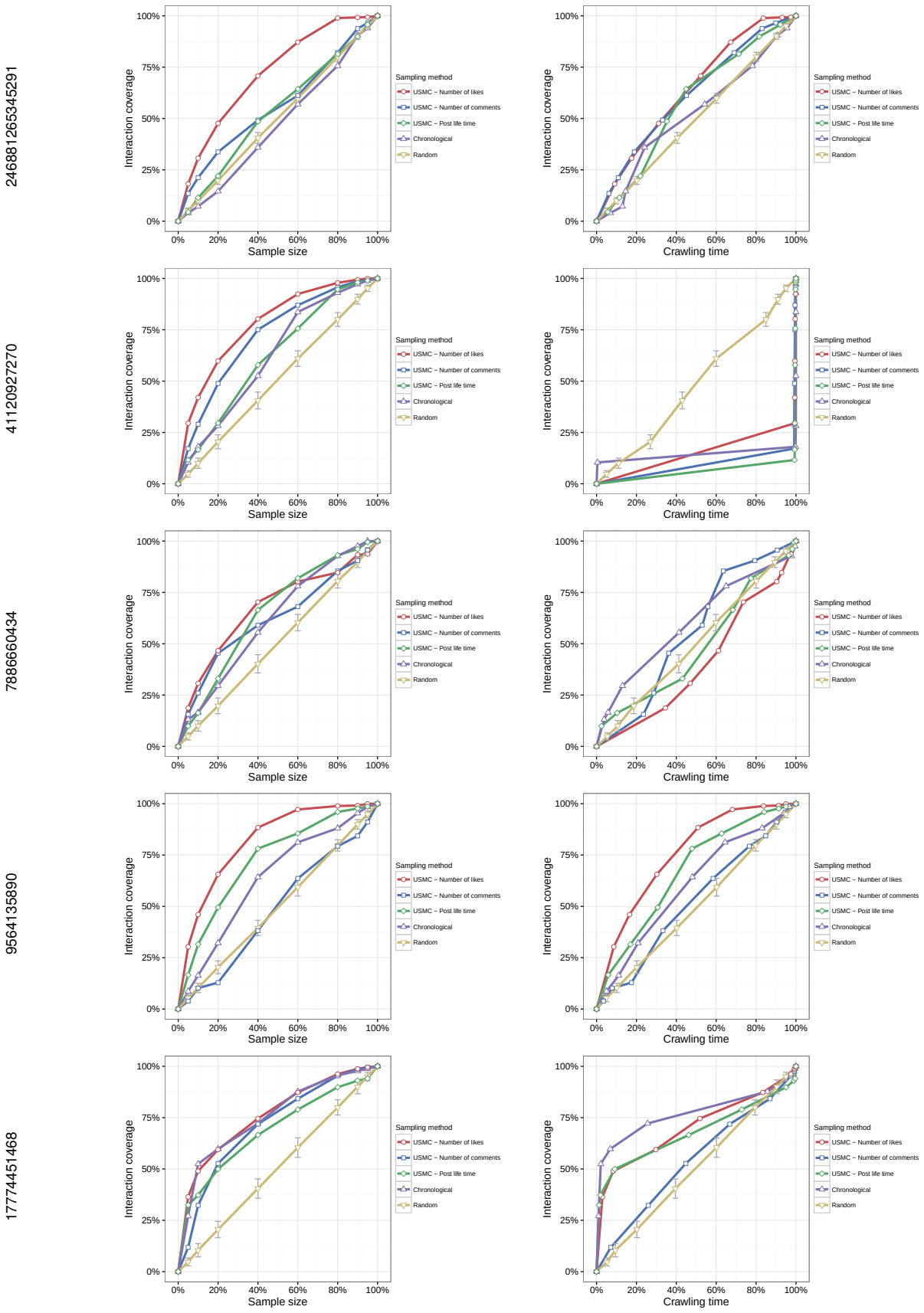
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



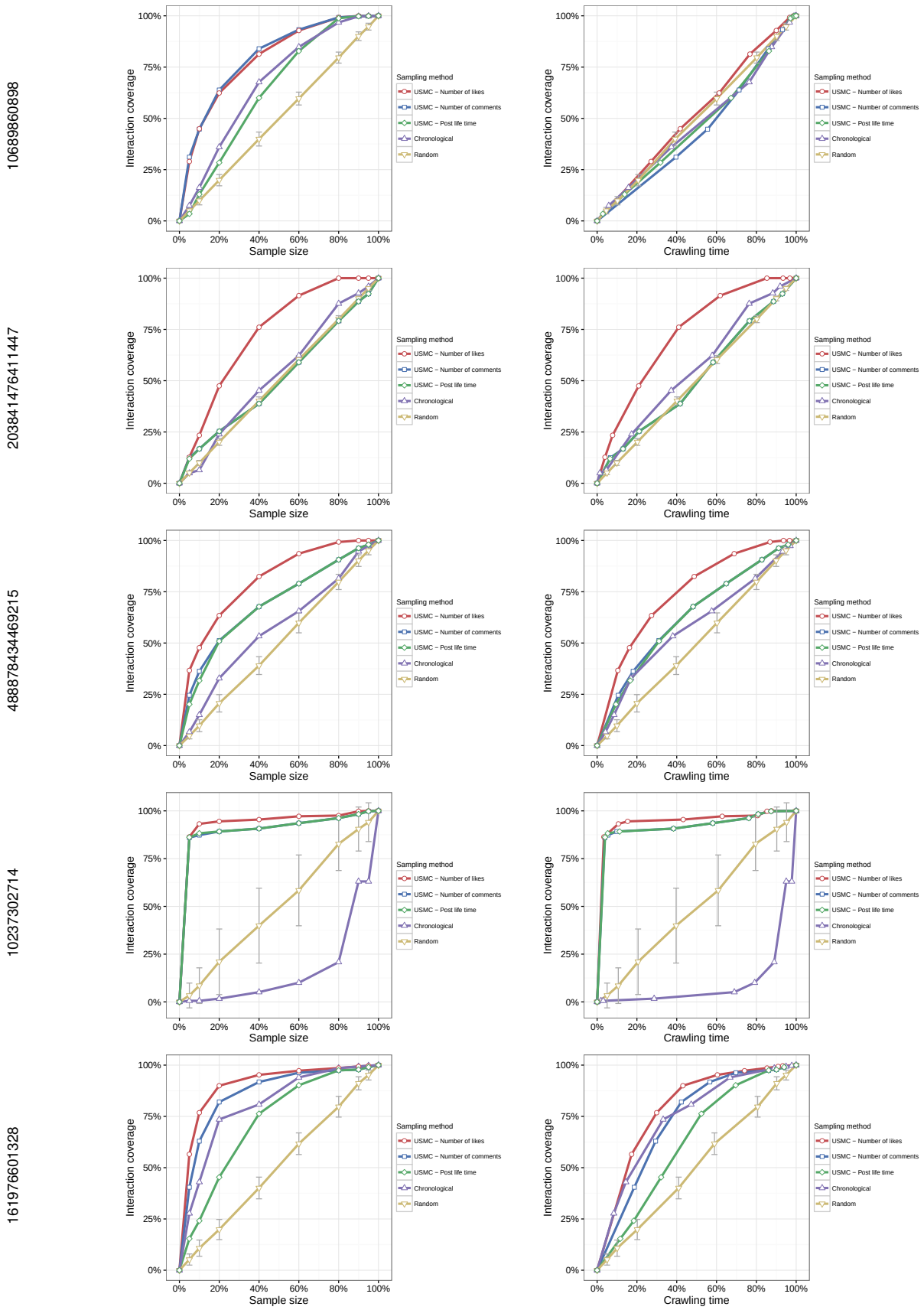
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



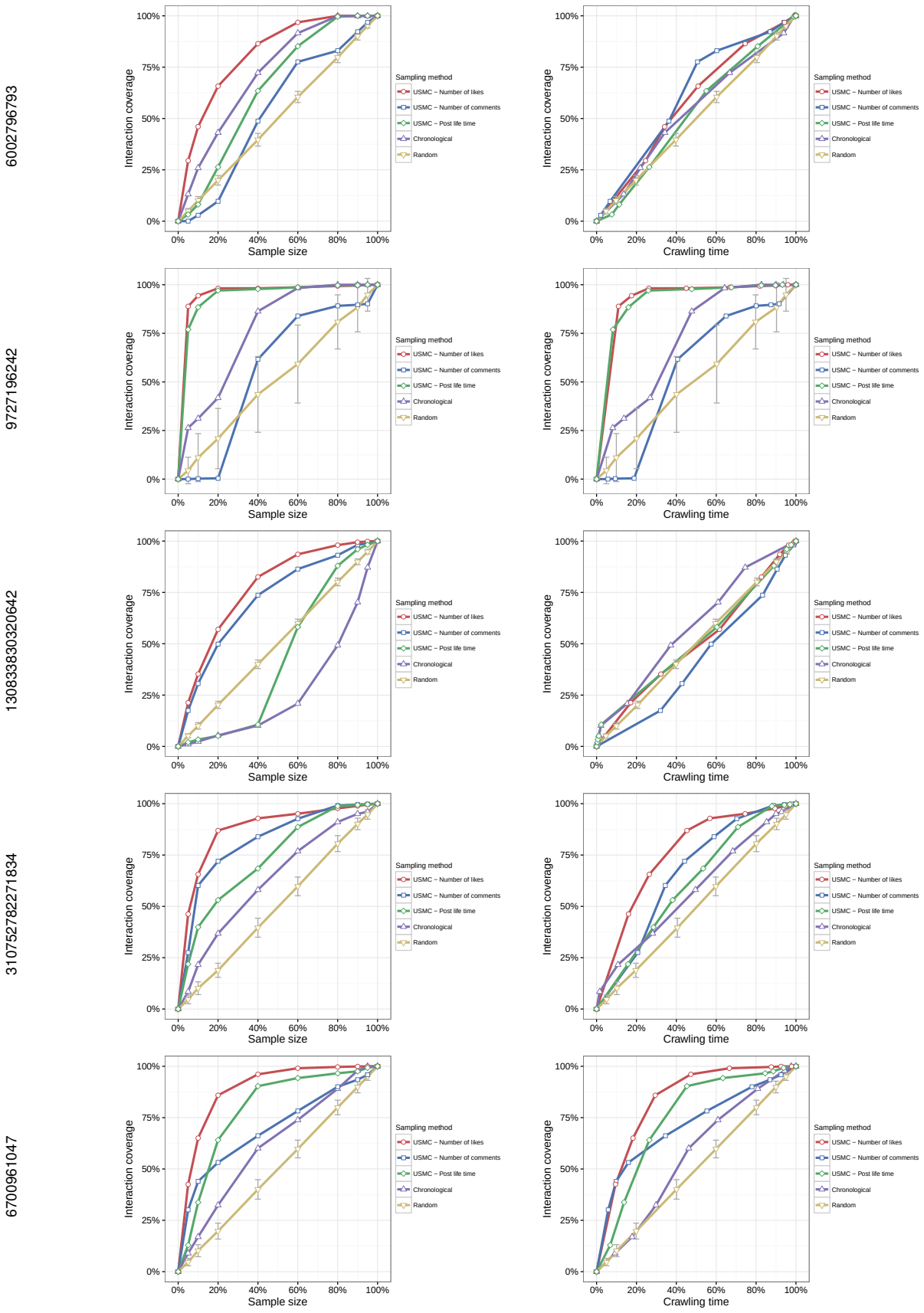
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



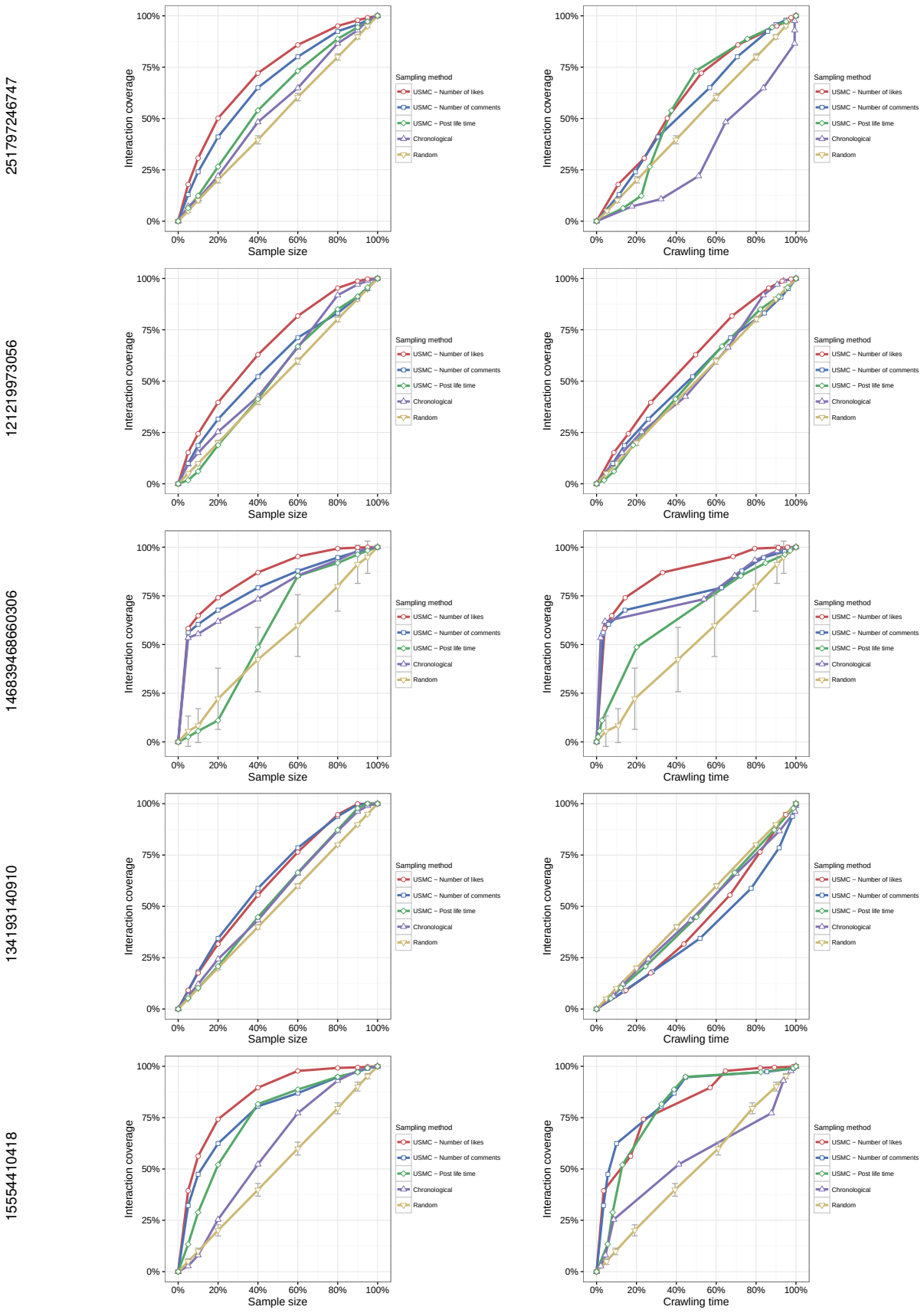
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



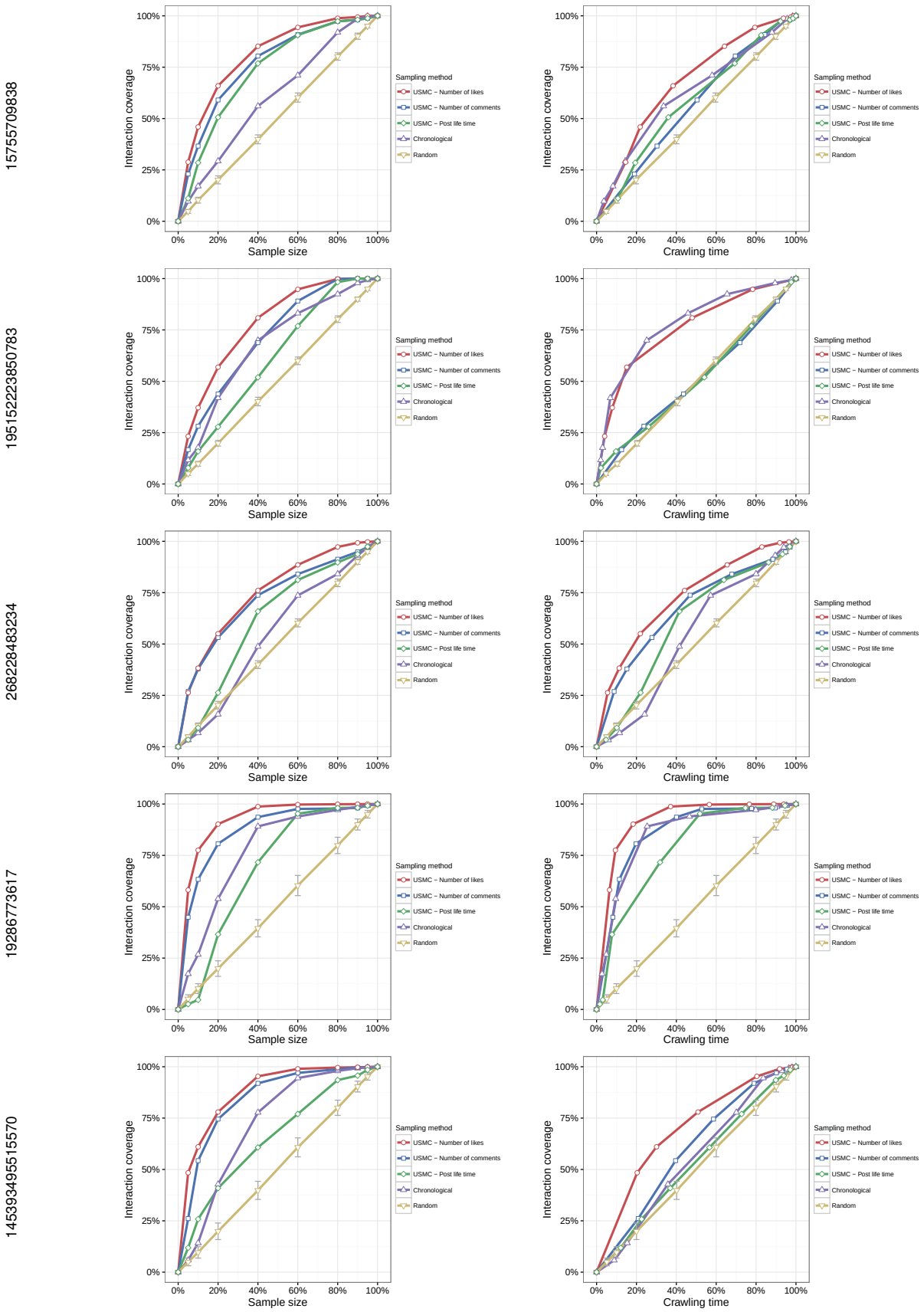
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



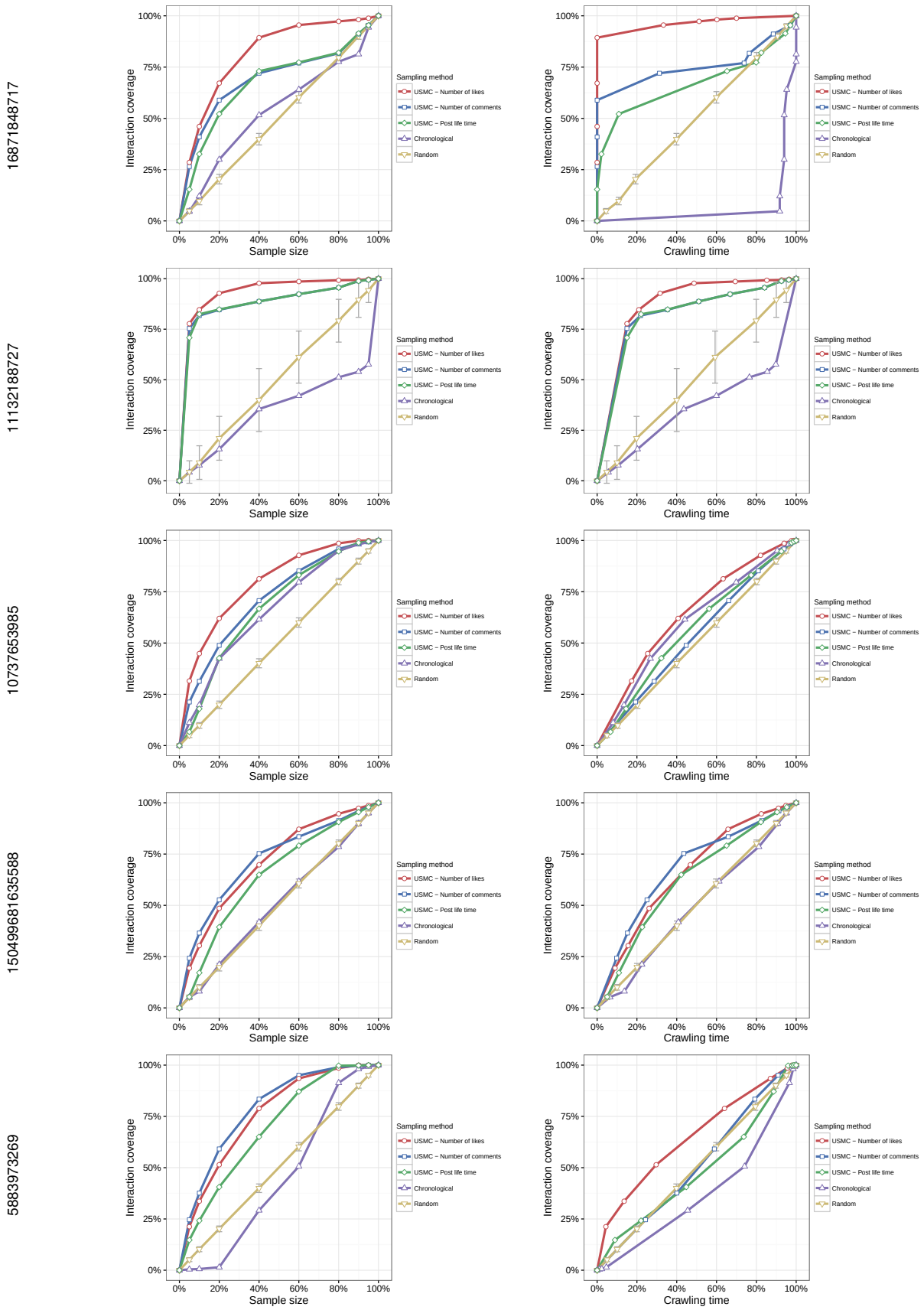
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



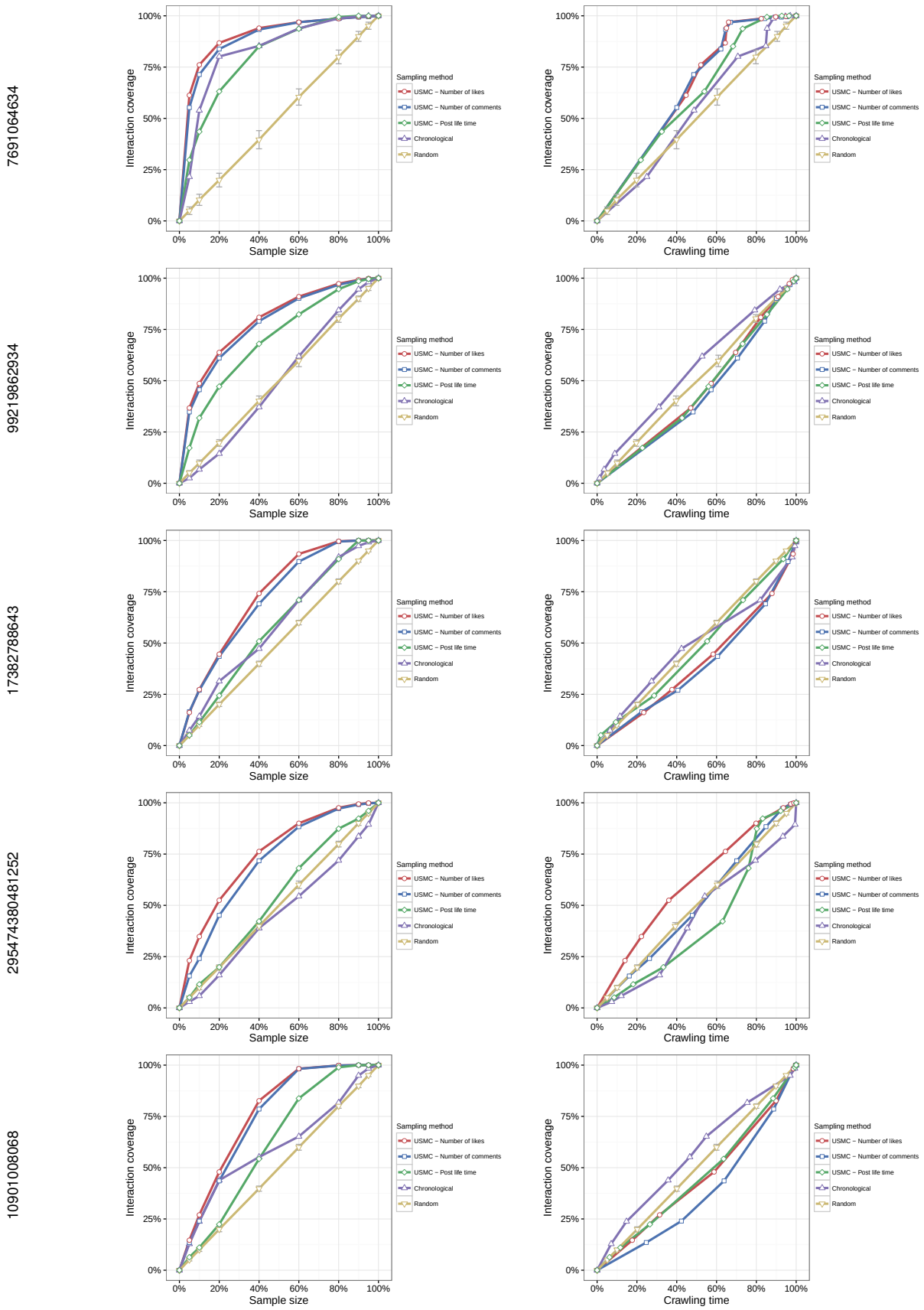
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



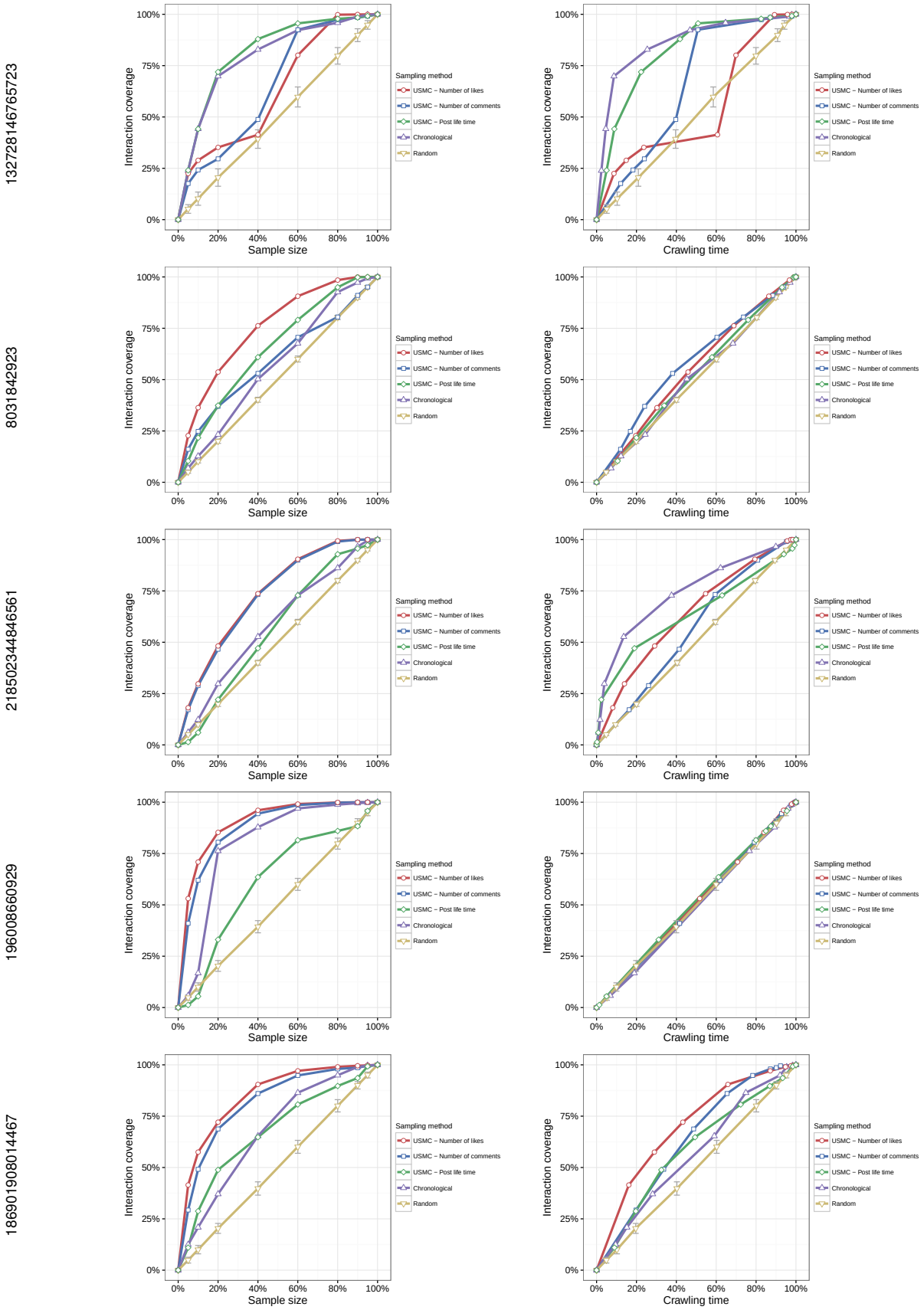
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



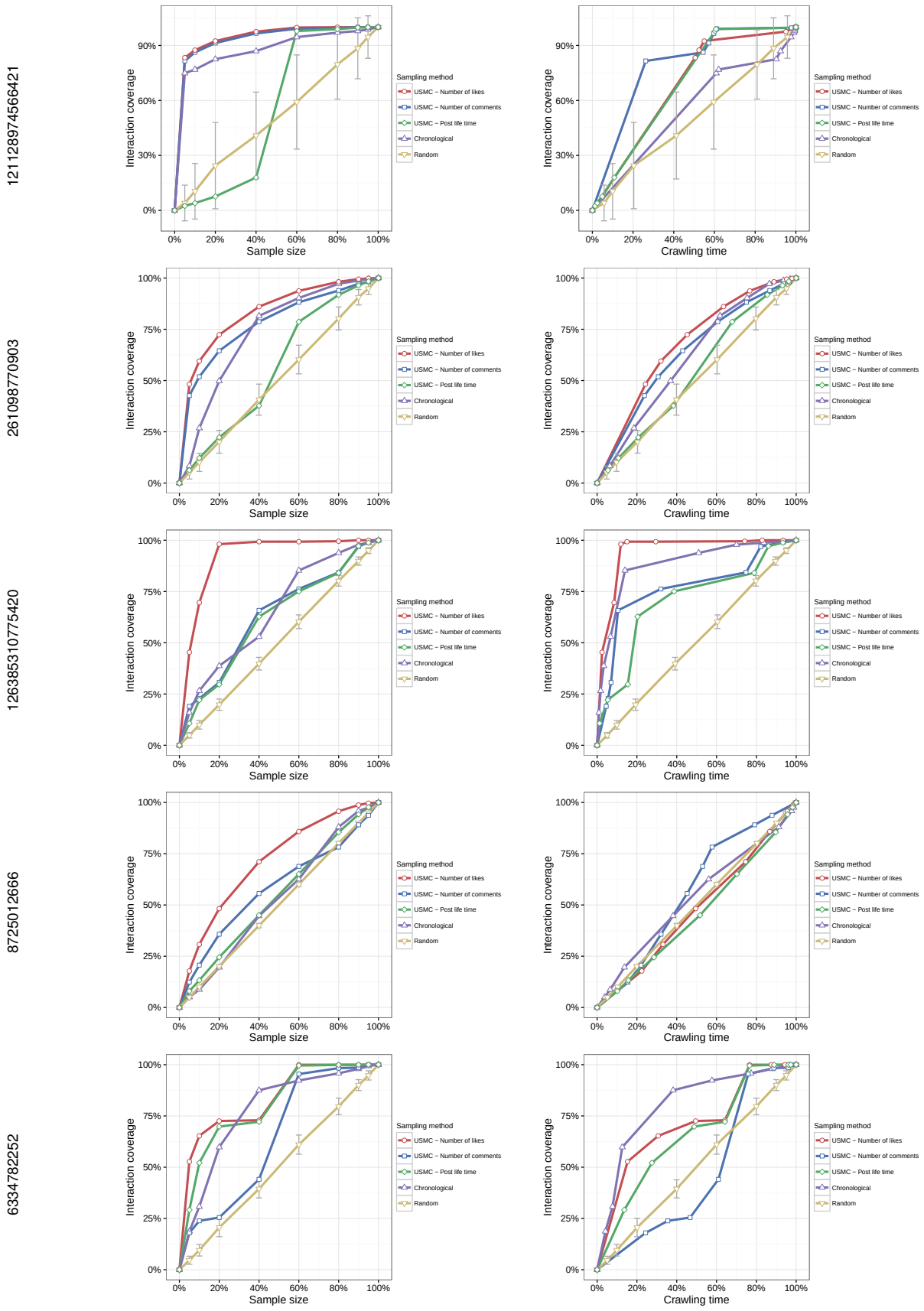
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



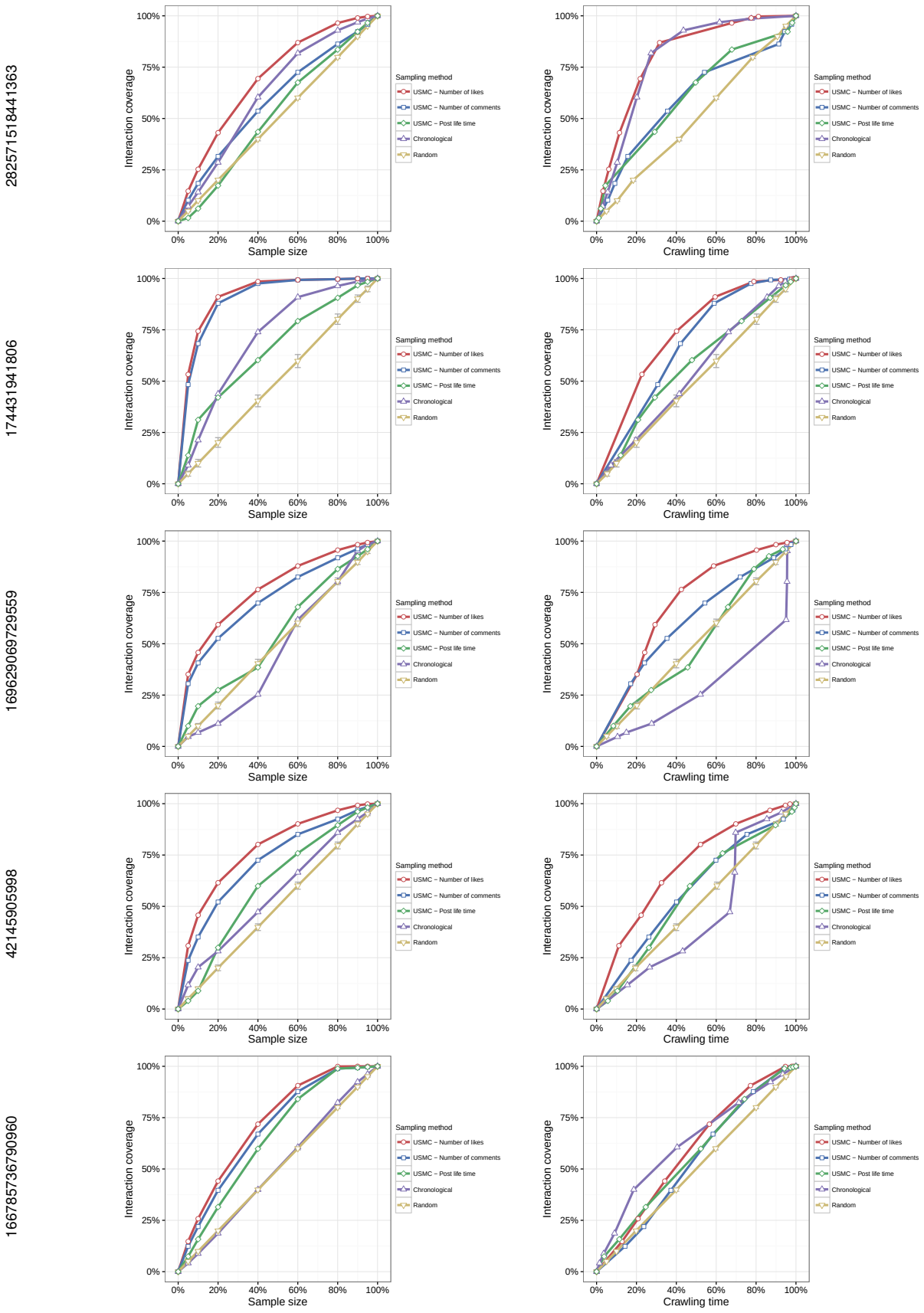
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



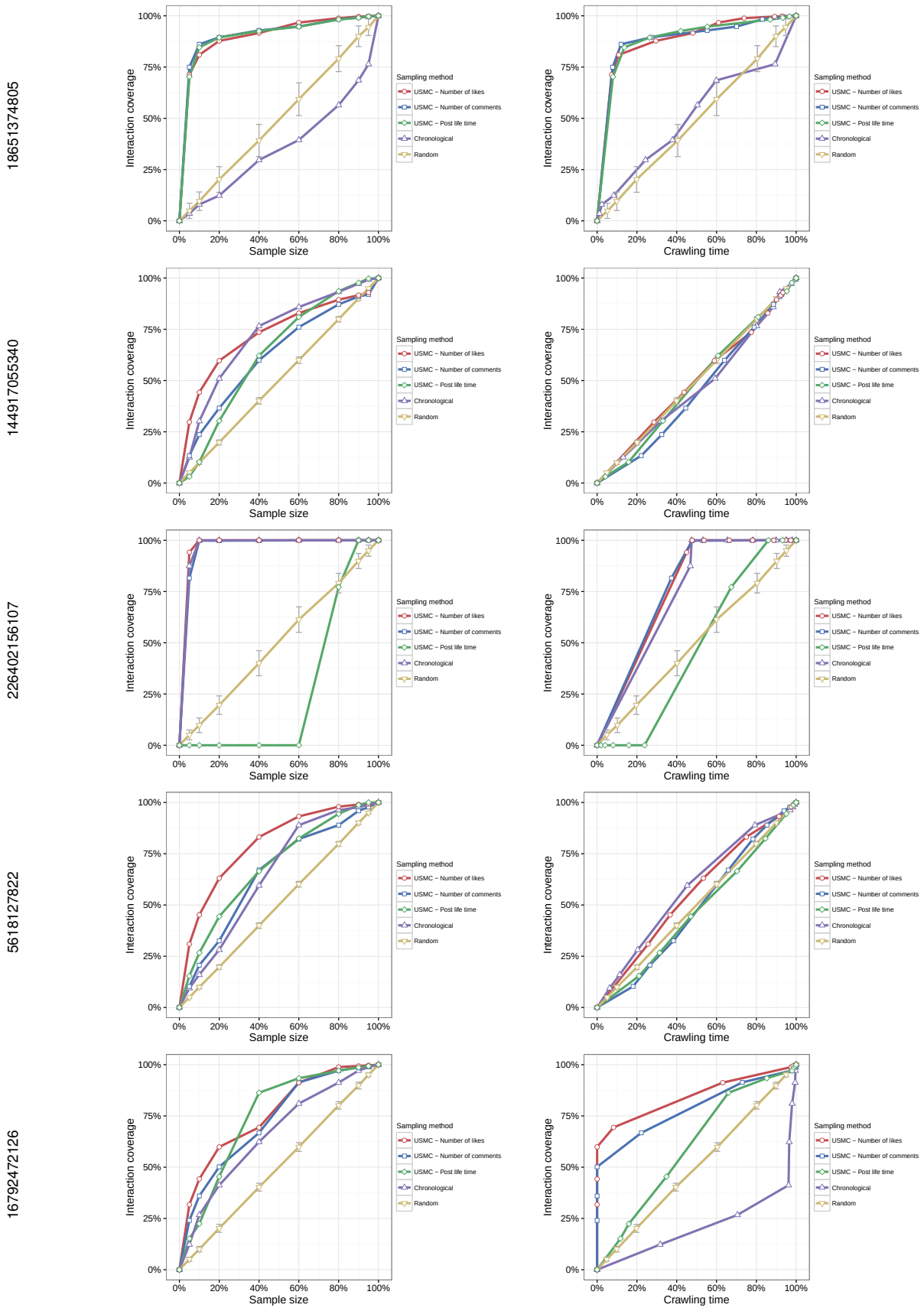
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



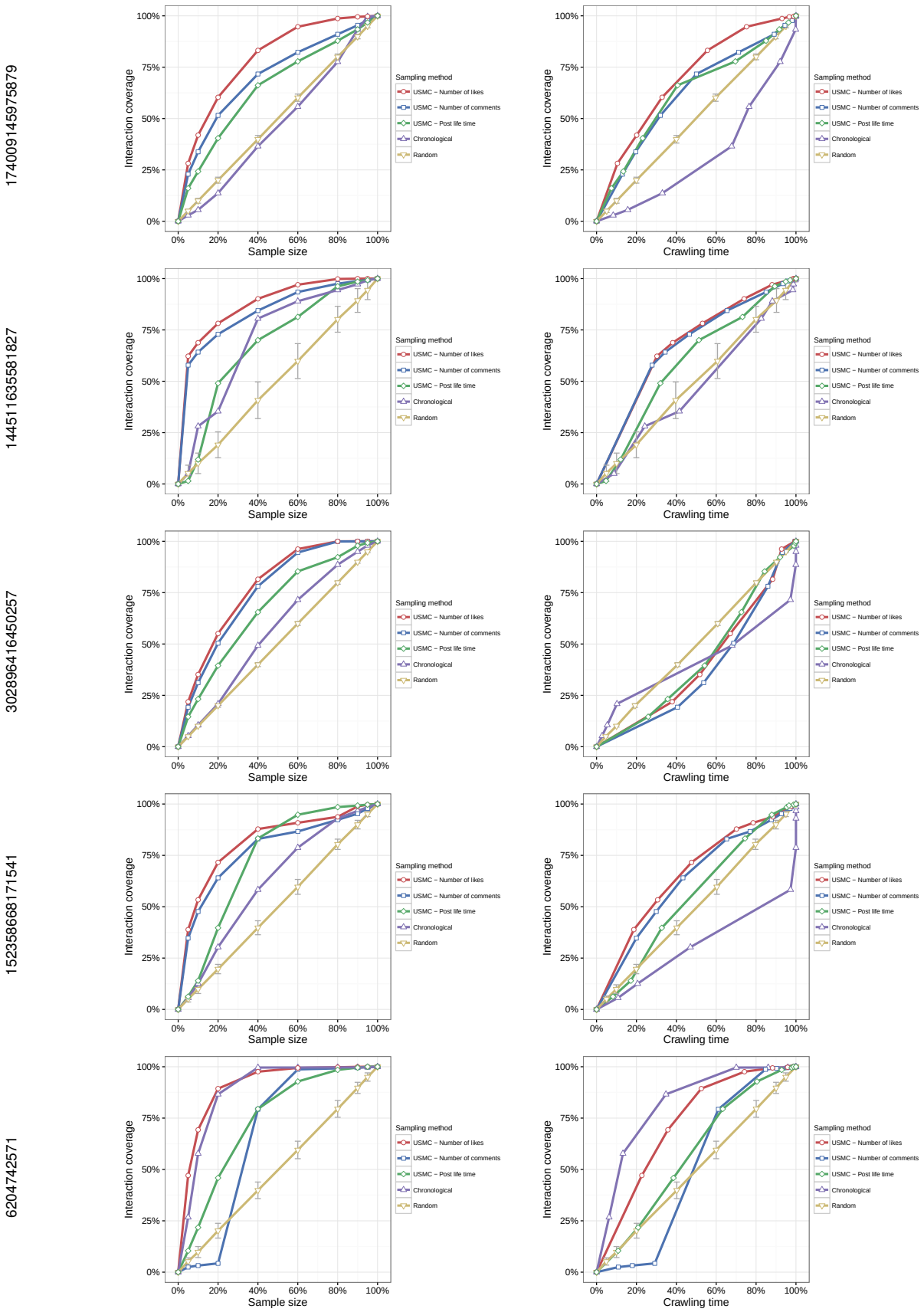
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



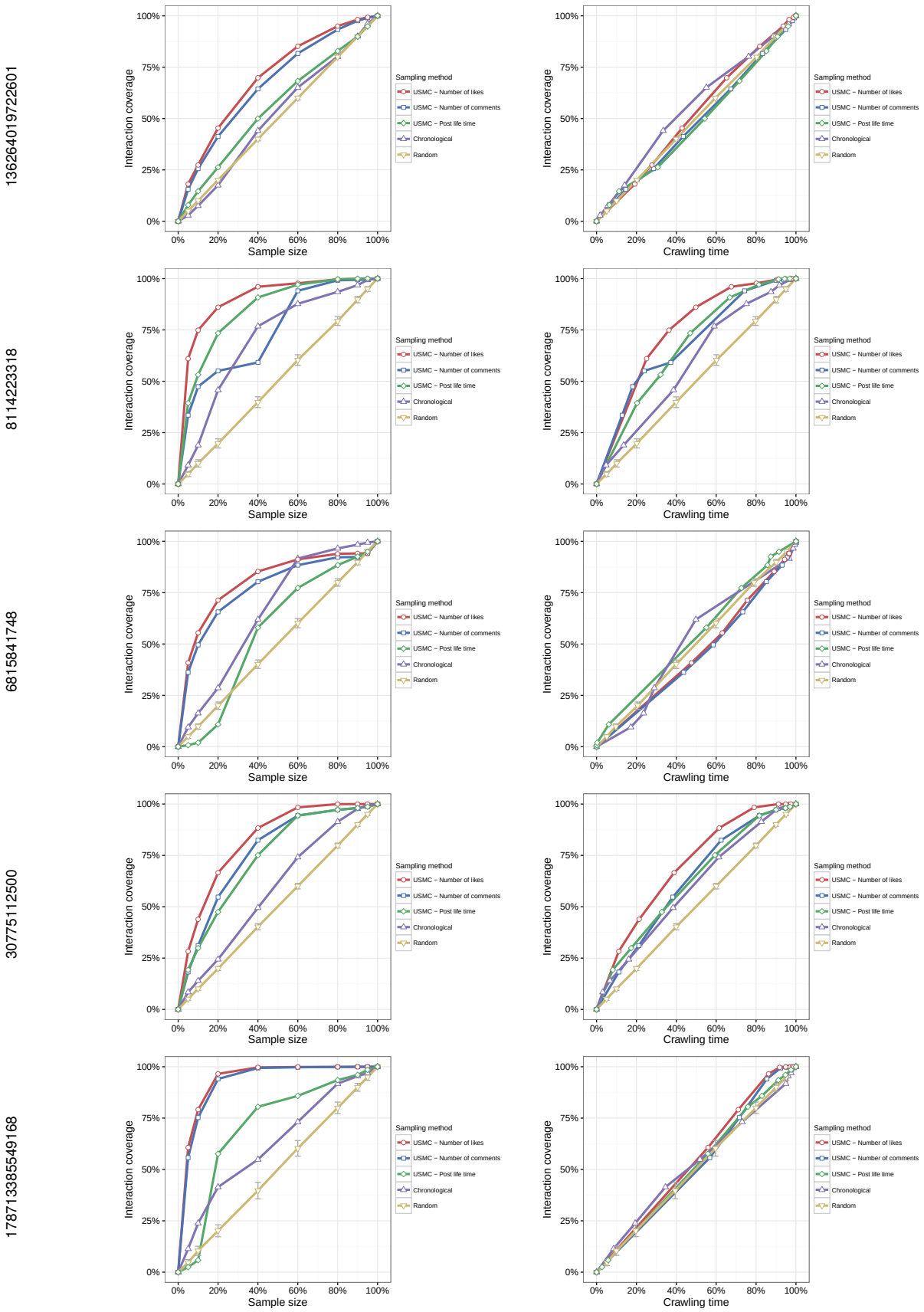
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



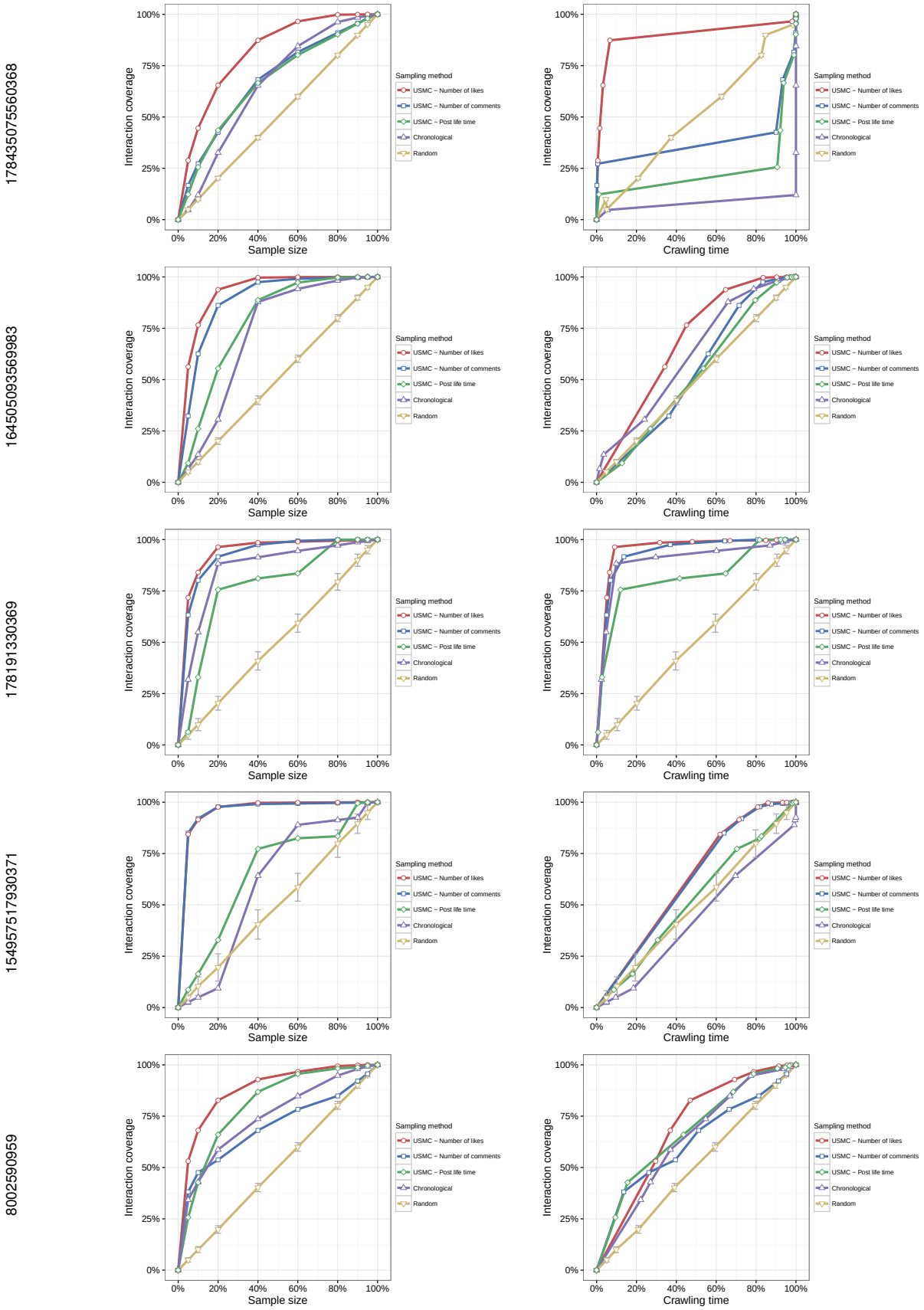
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



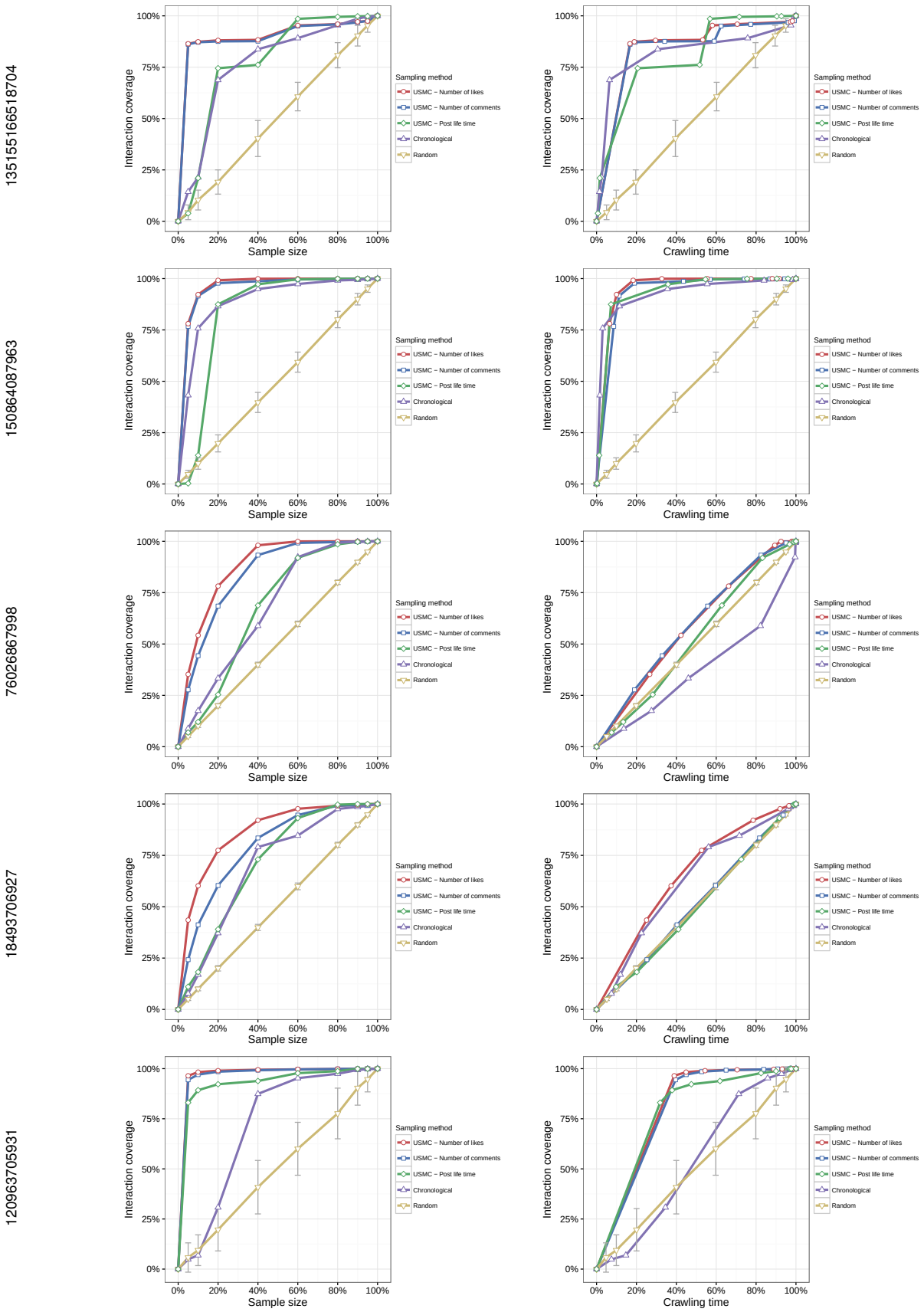
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



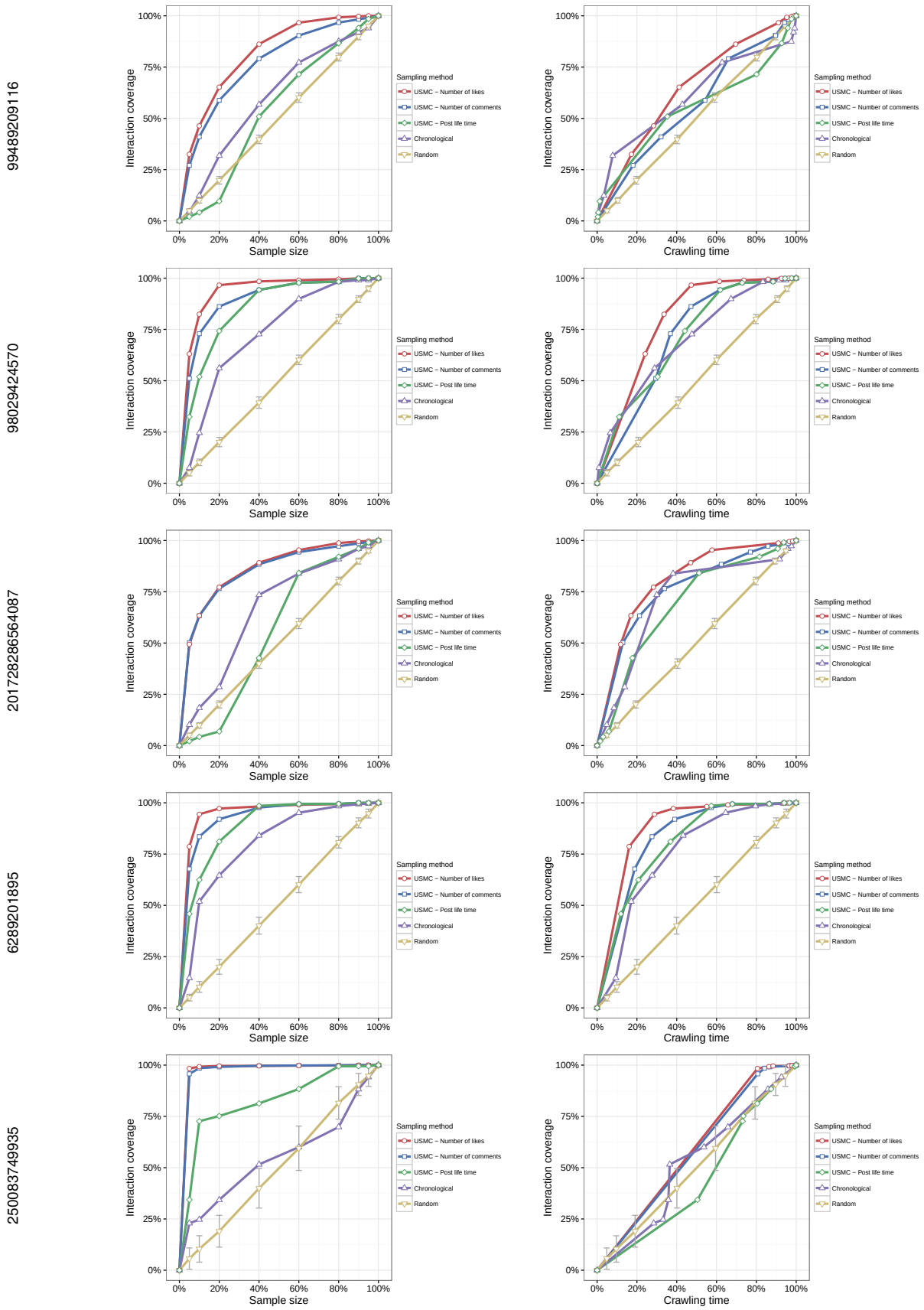
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



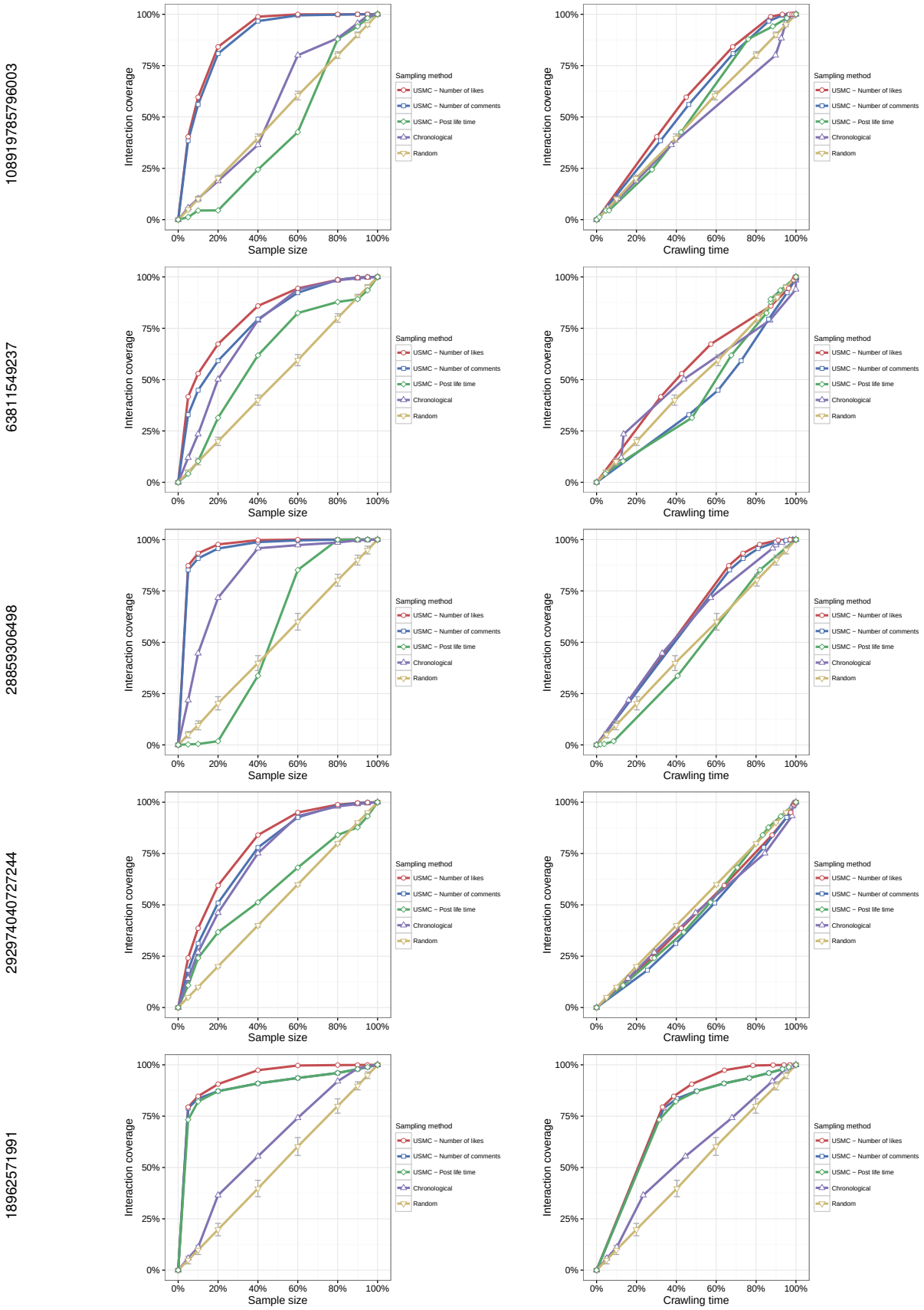
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



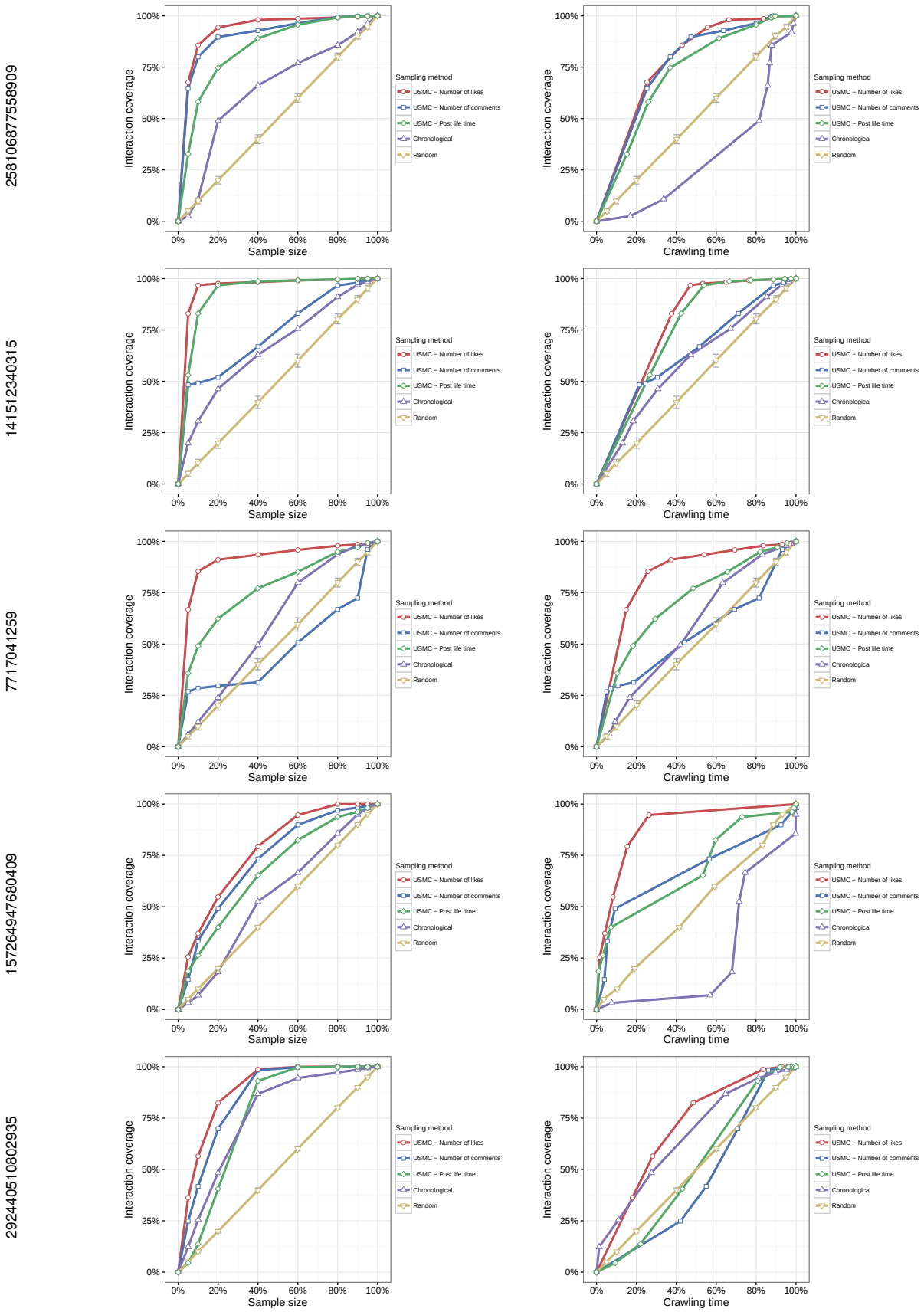
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



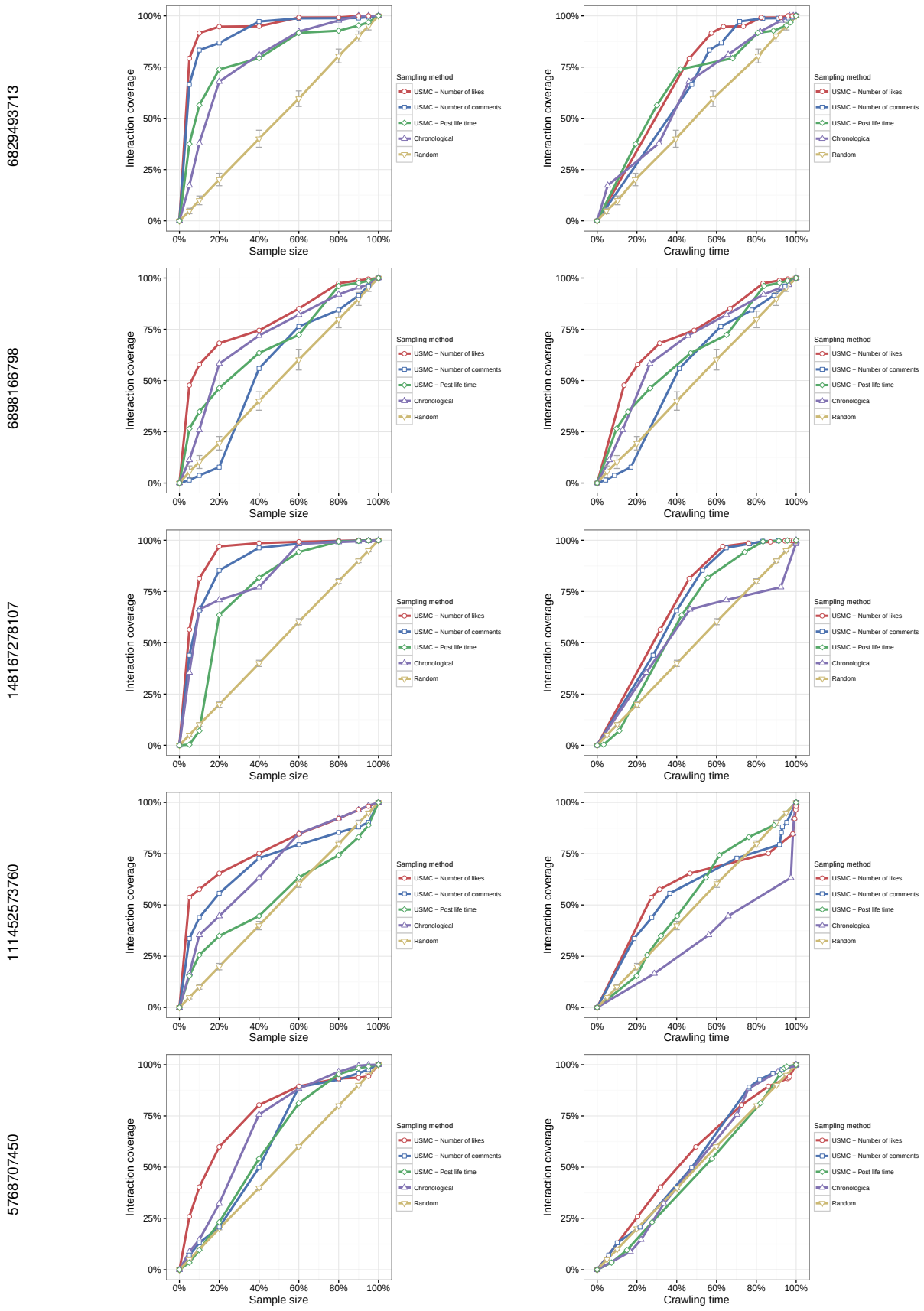
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



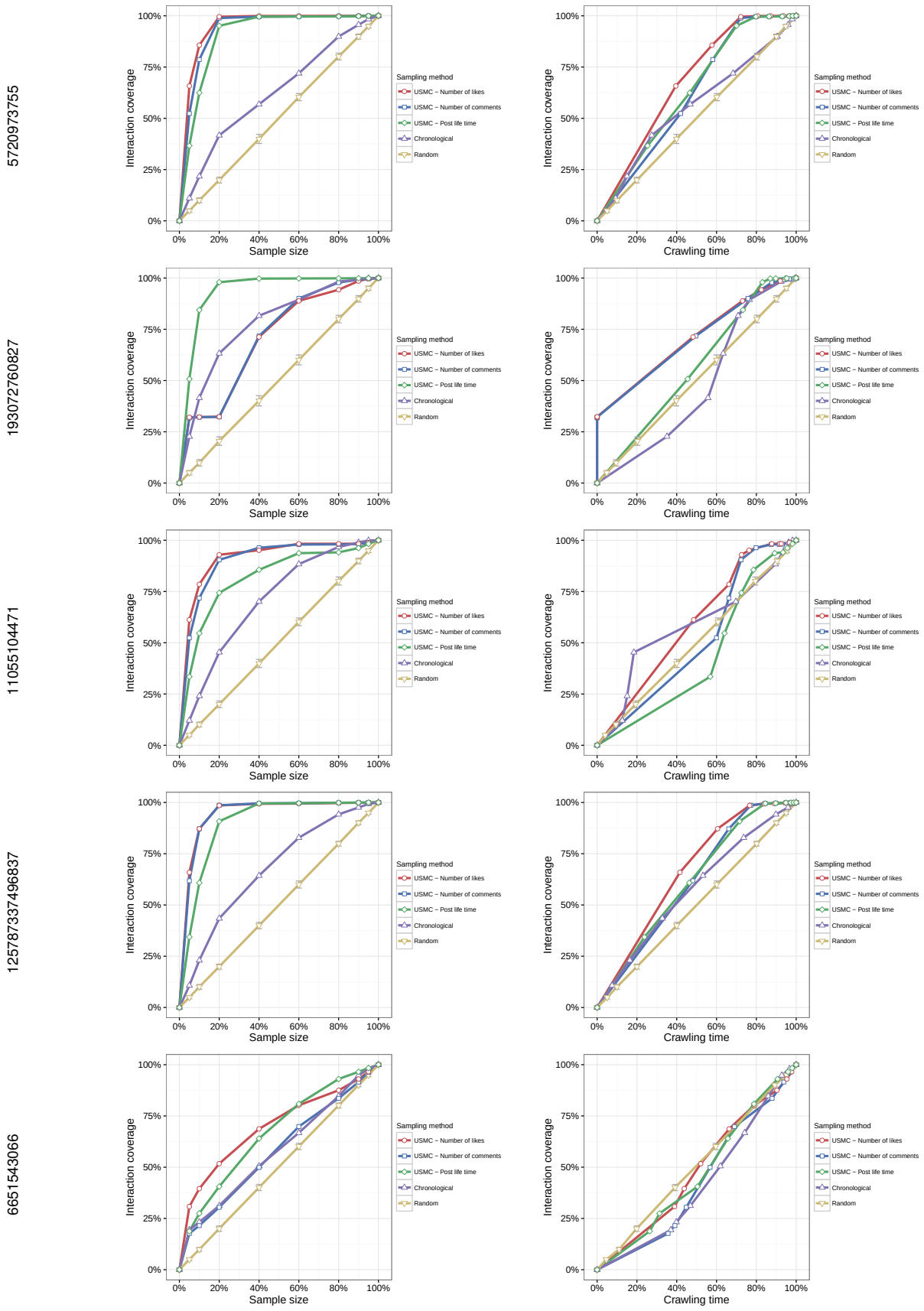
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



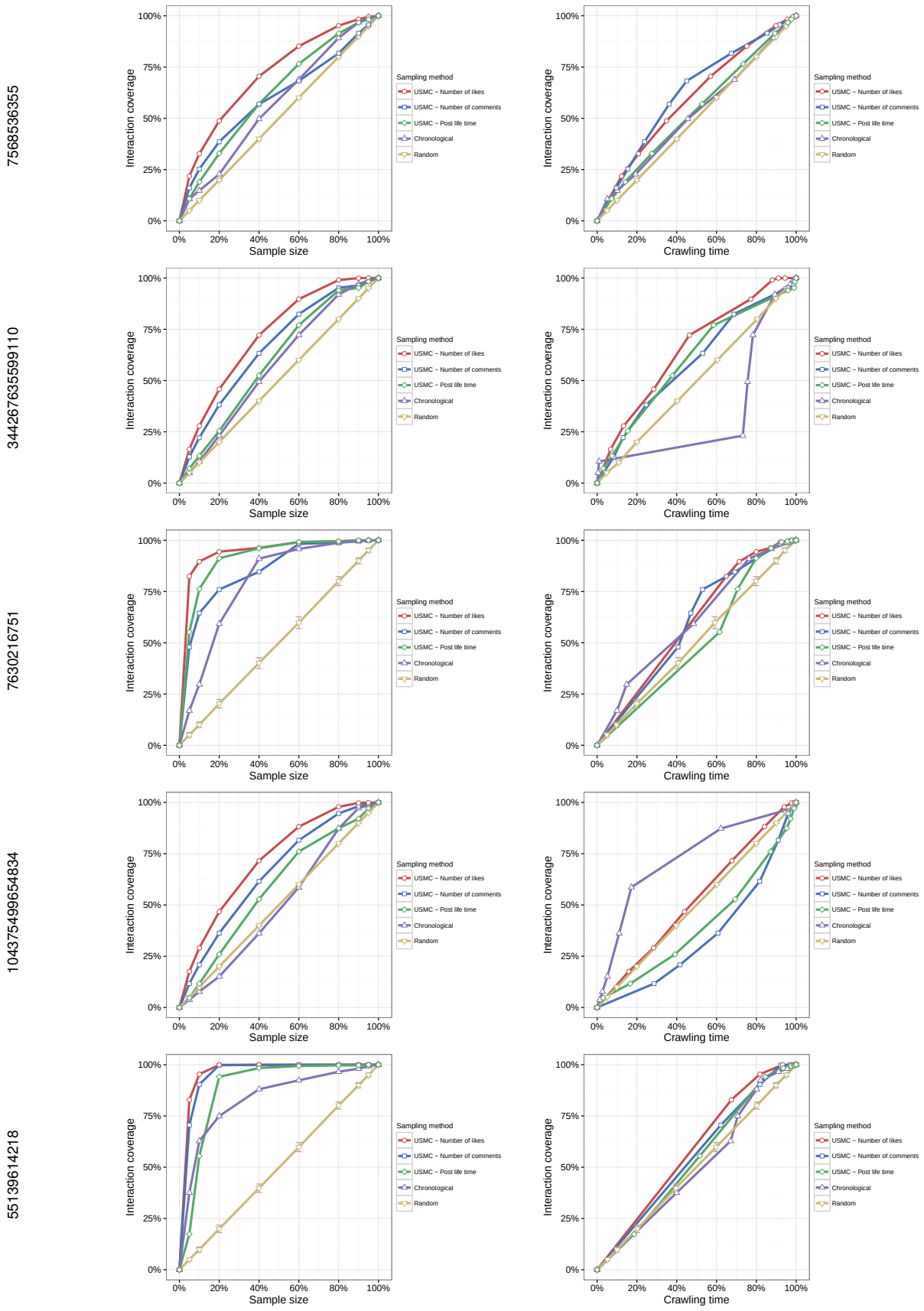
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



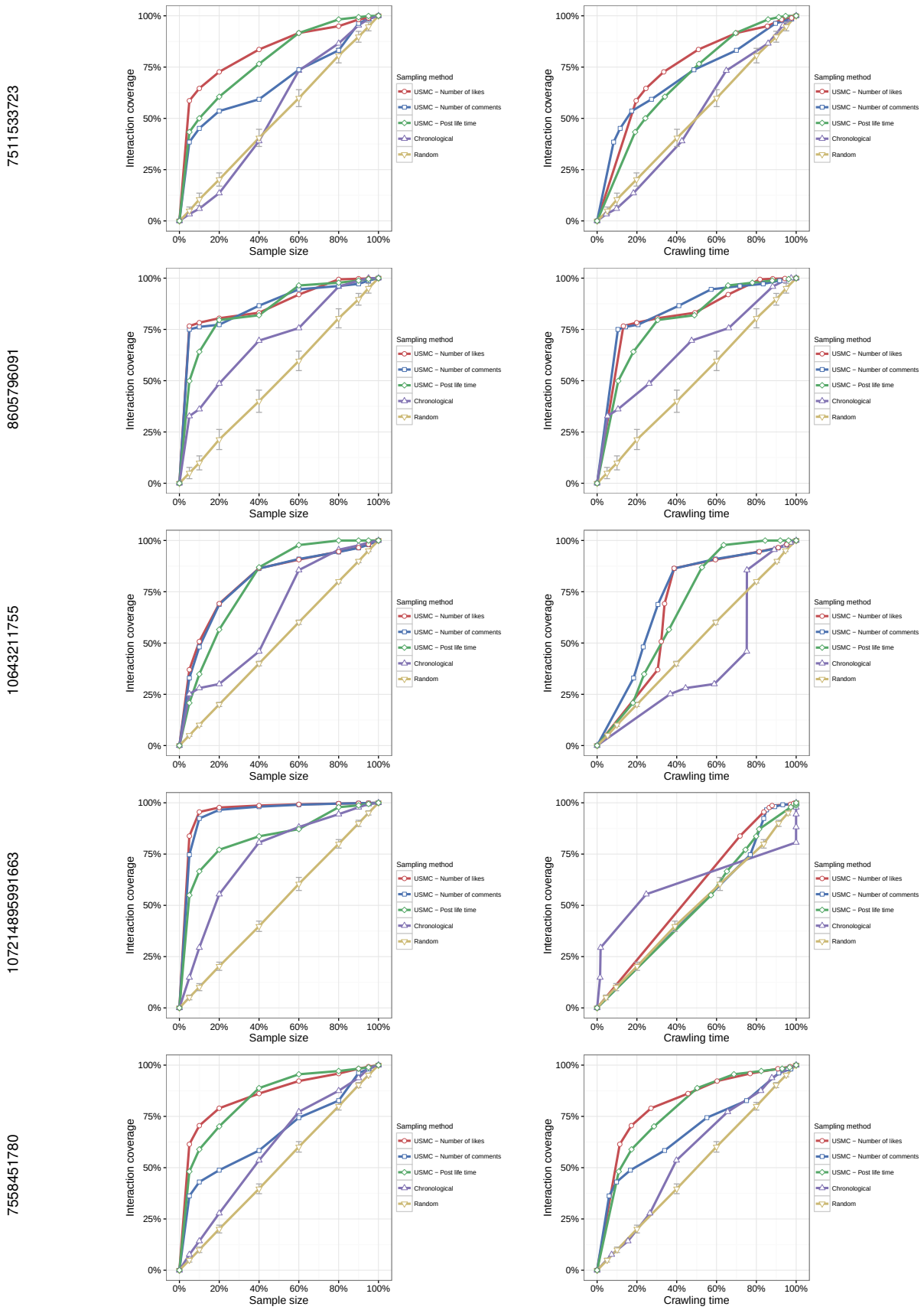
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



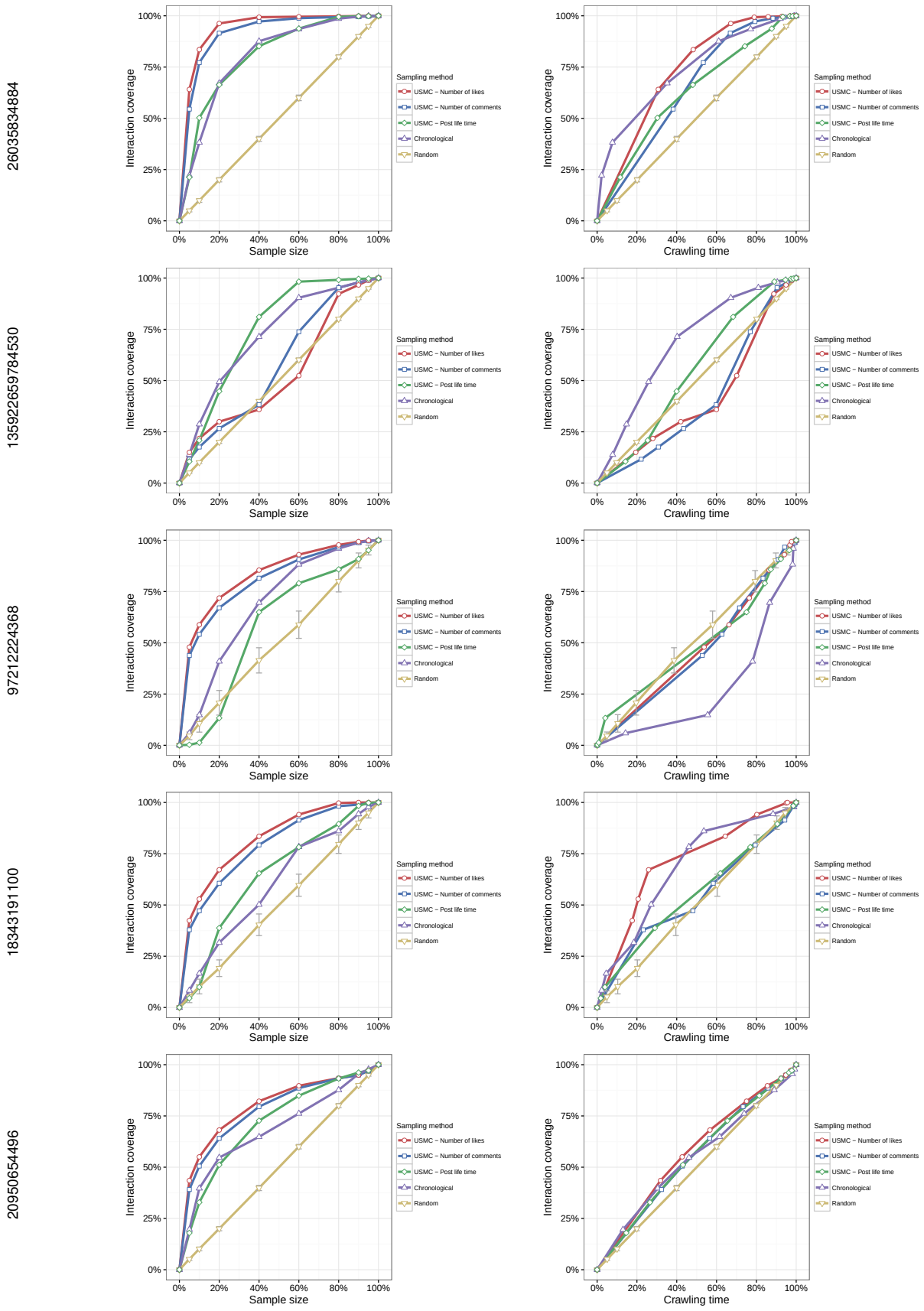
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



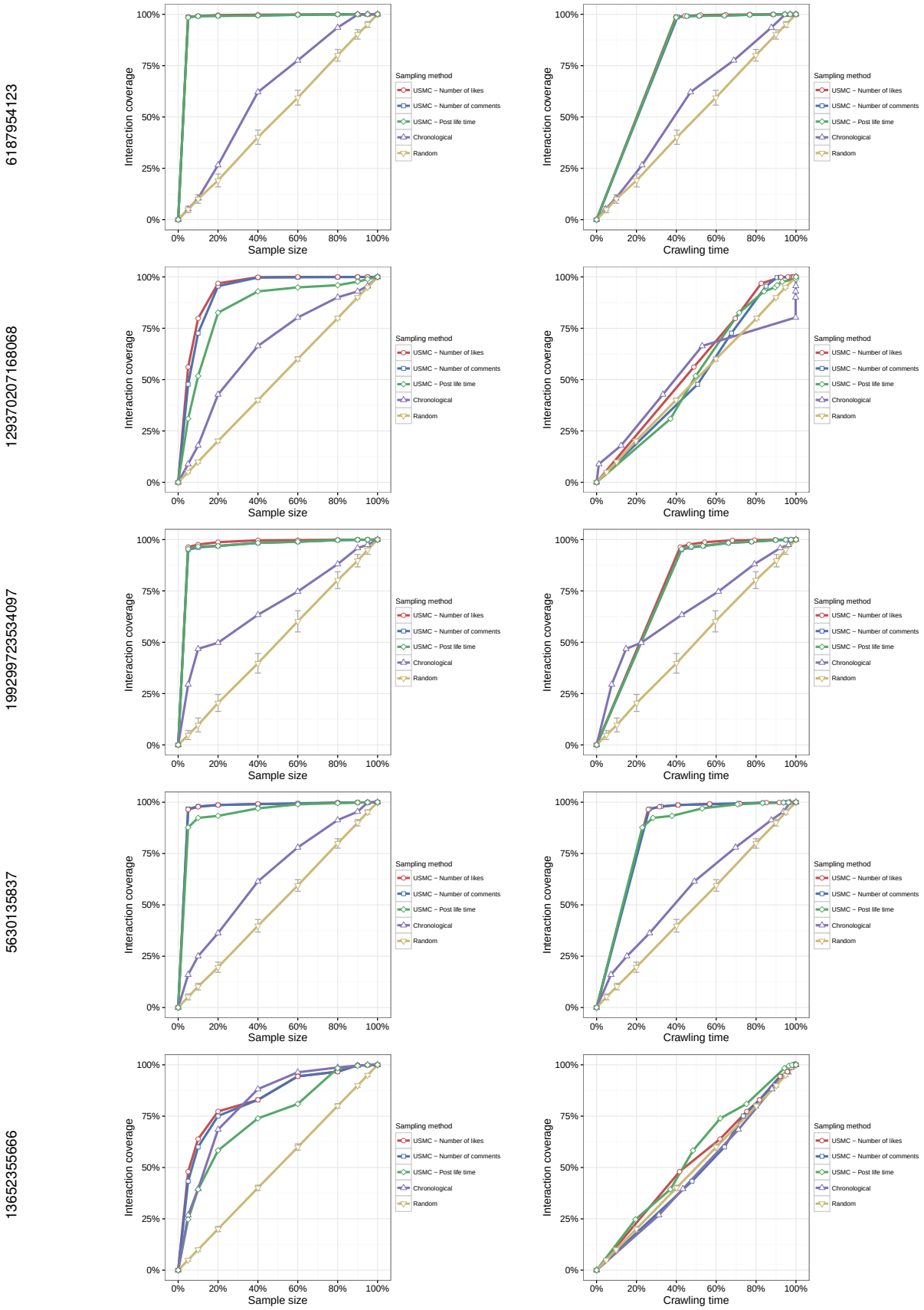
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



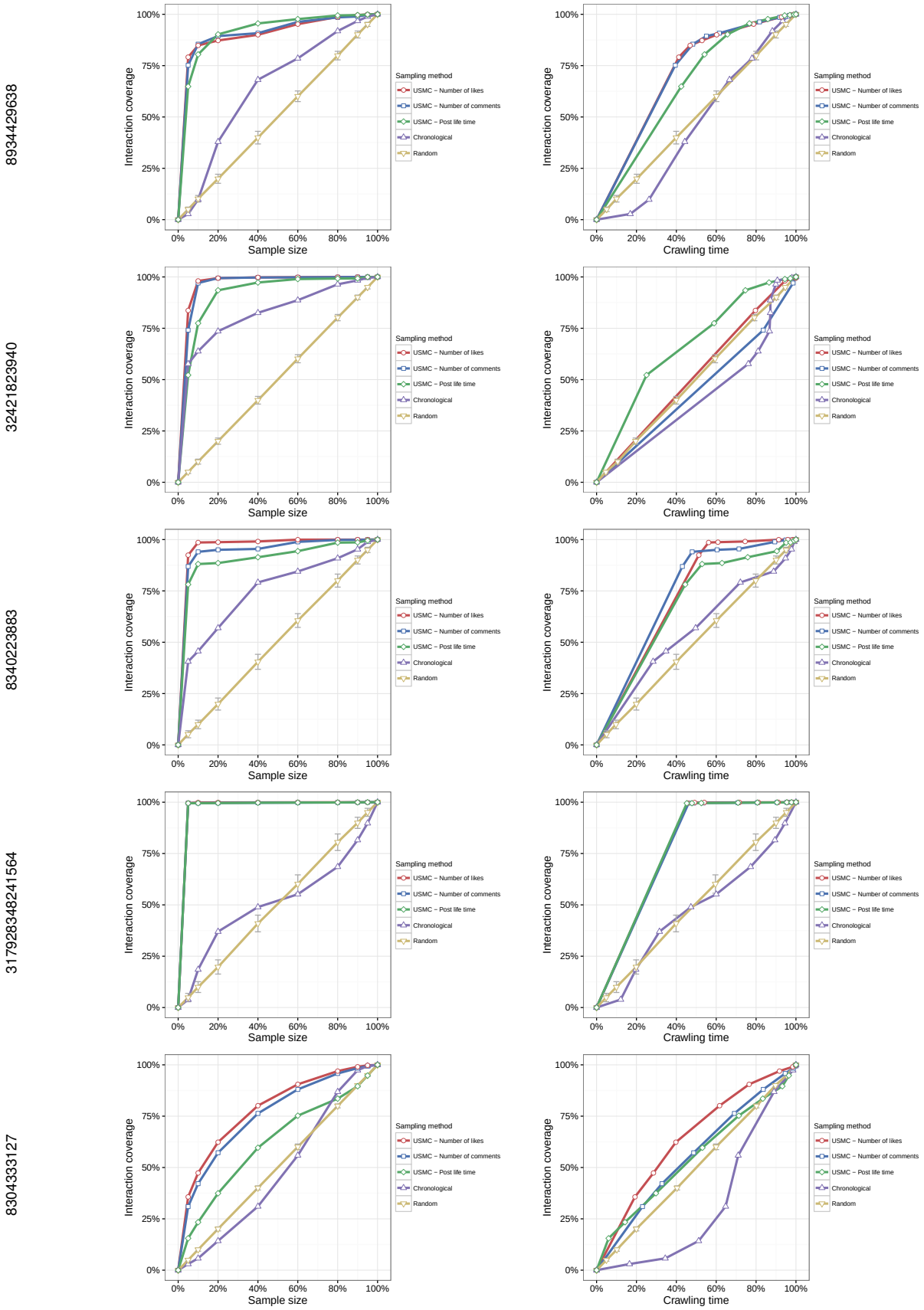
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



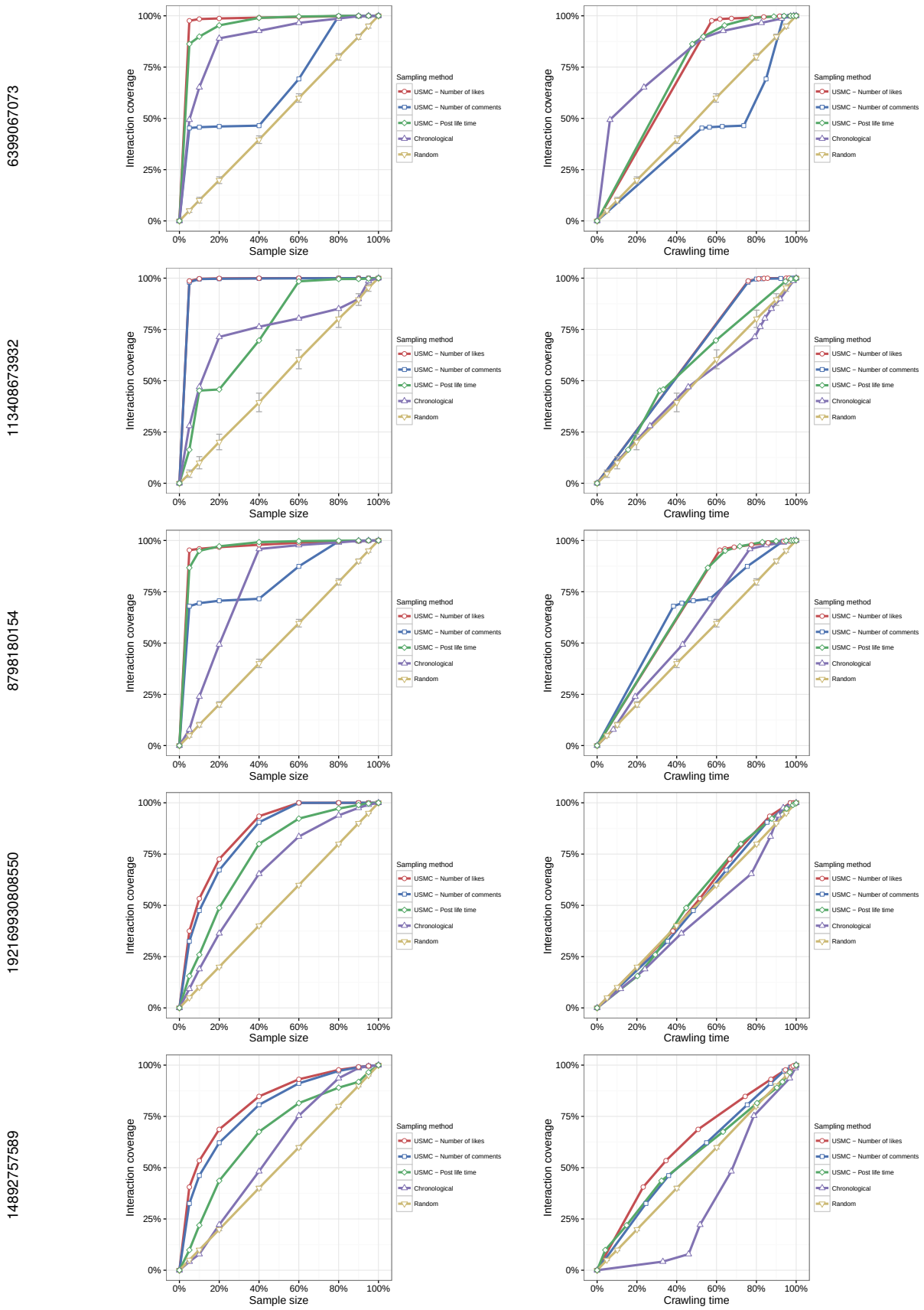
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



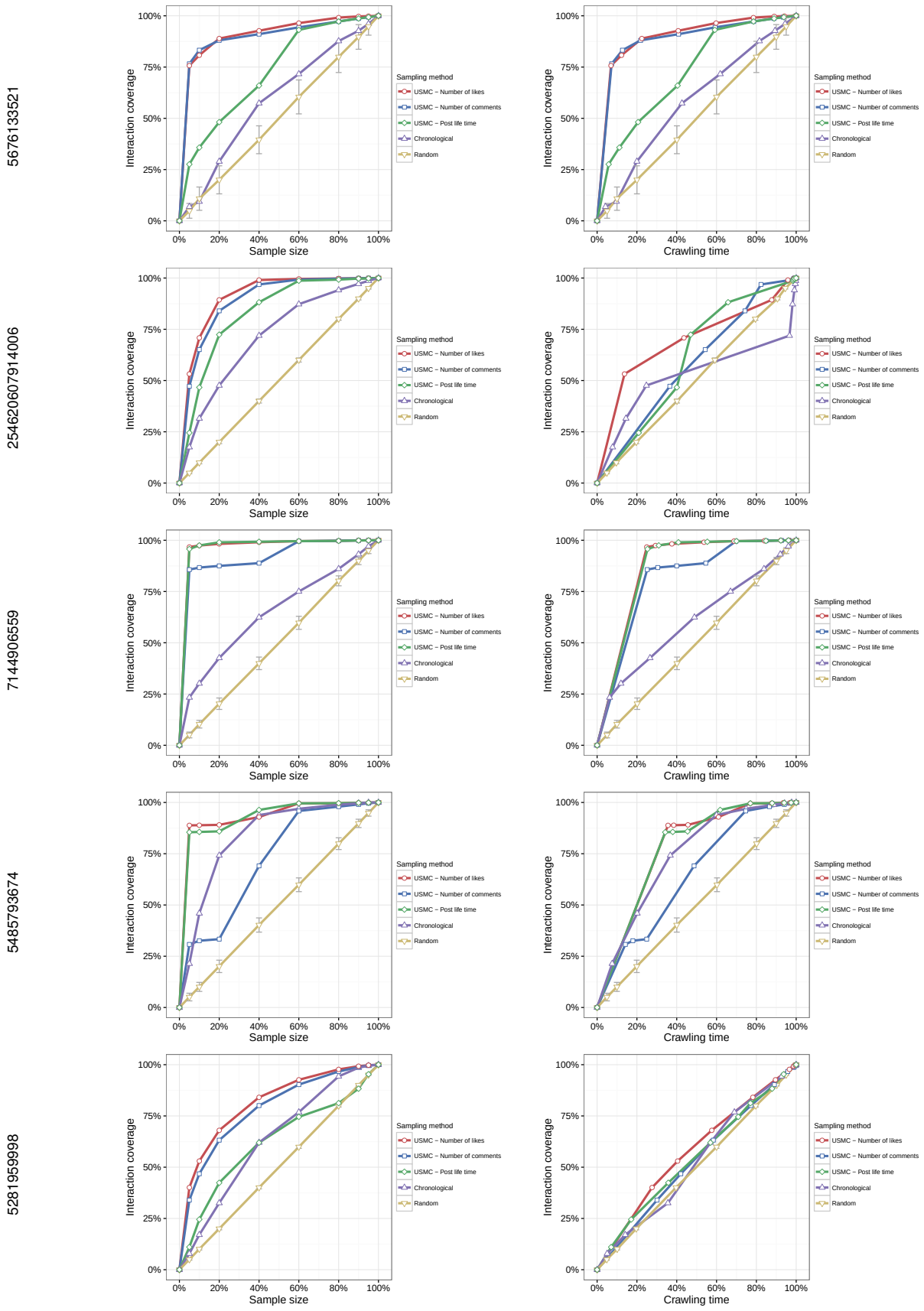
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



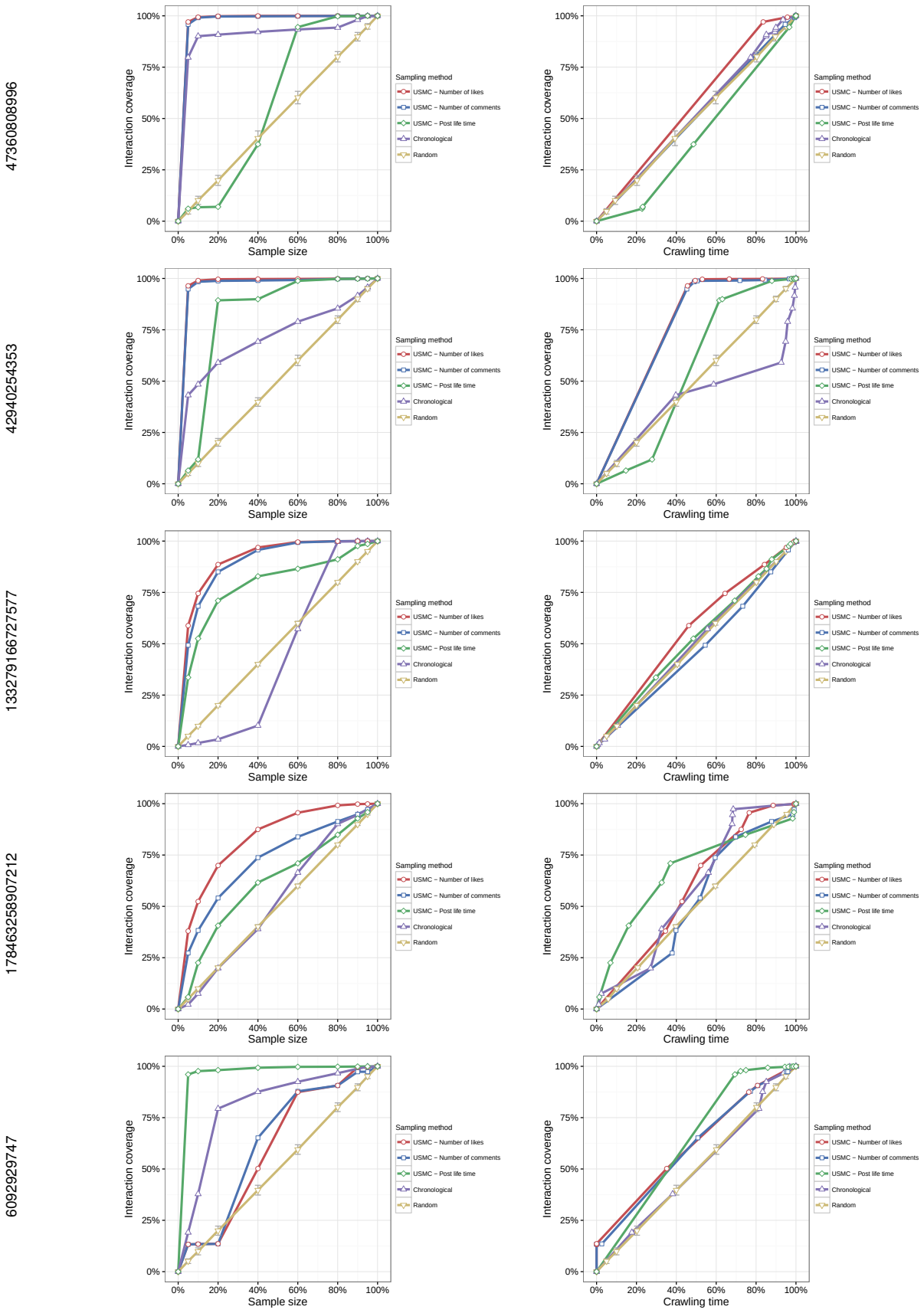
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



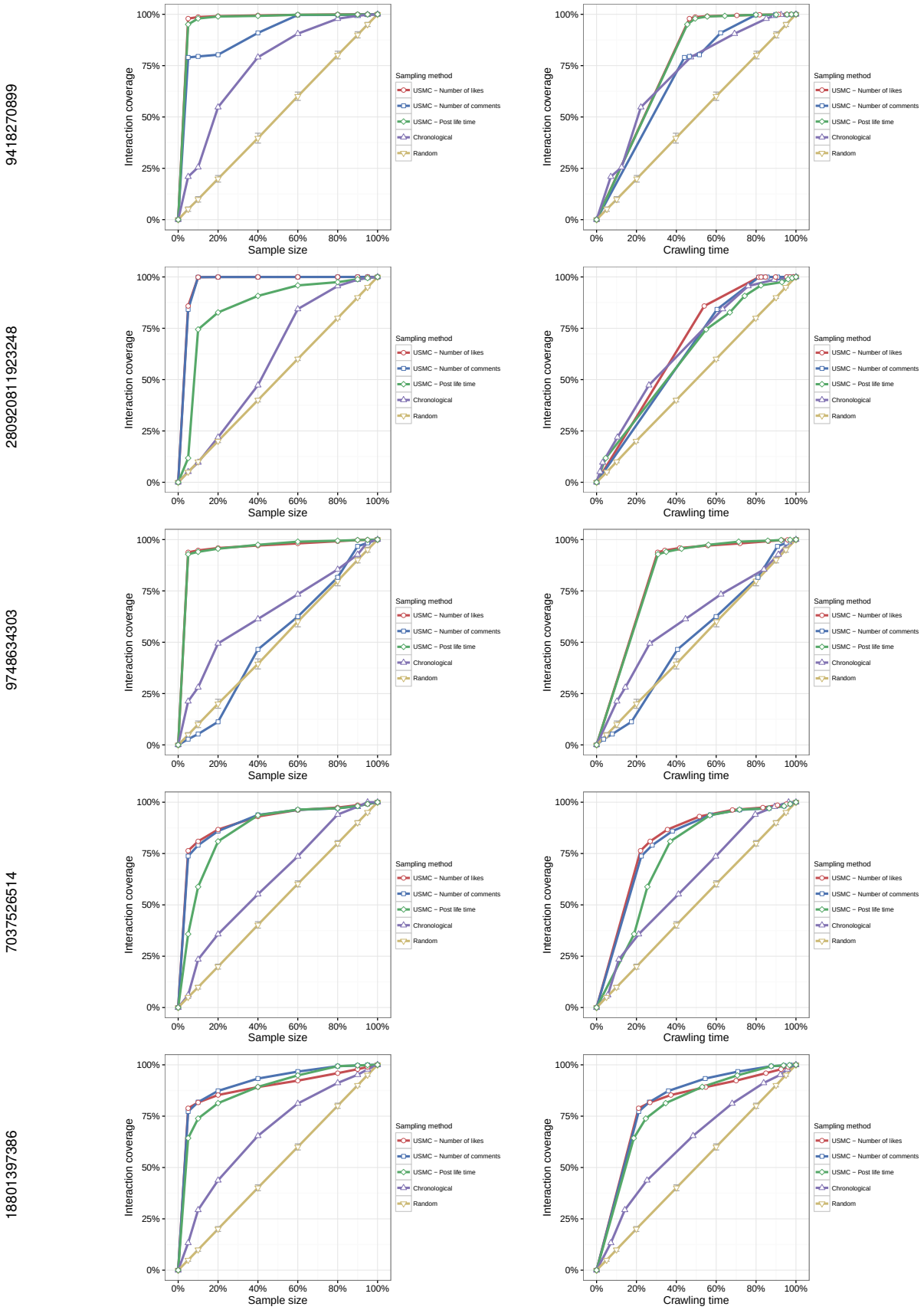
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



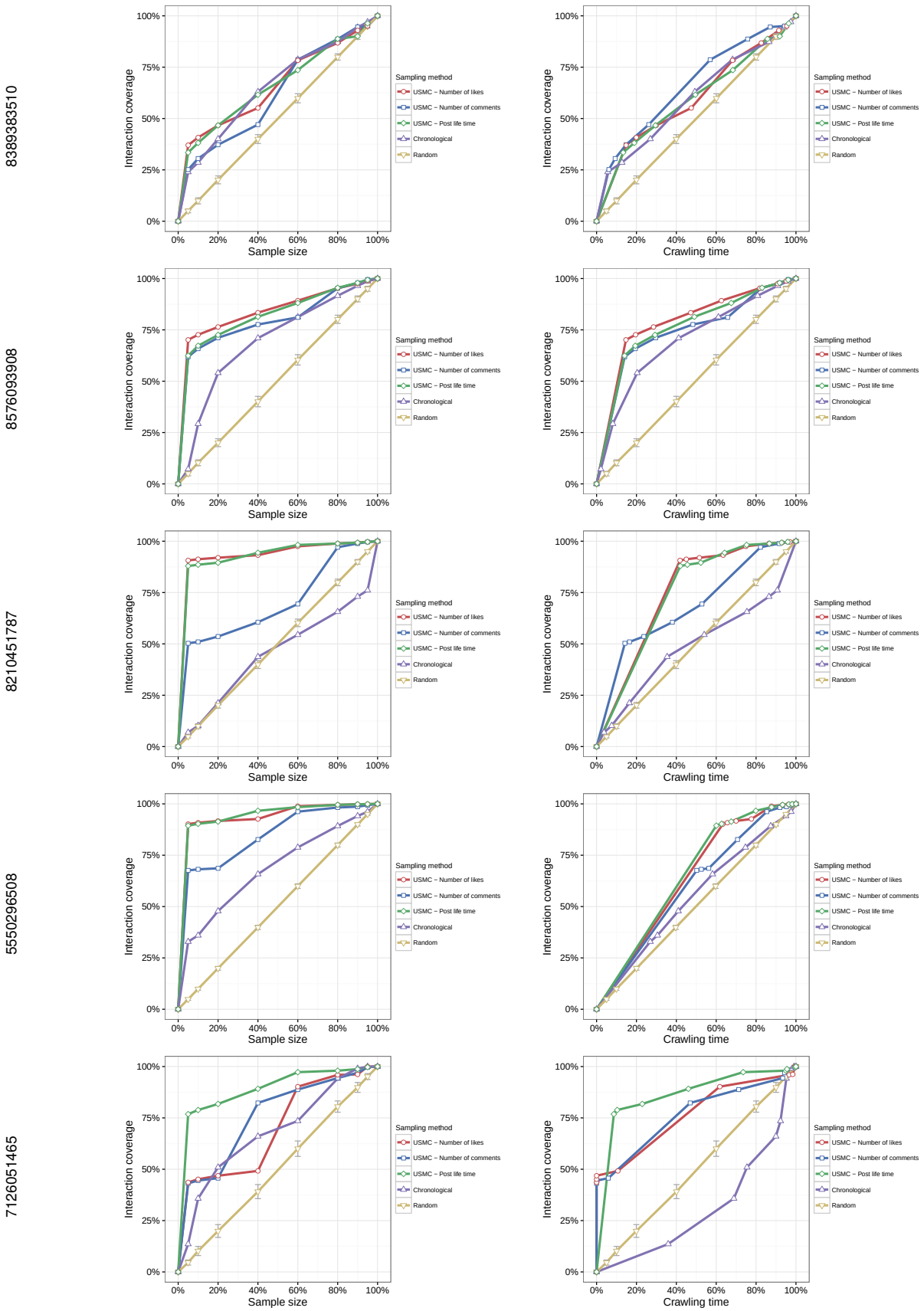
Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.

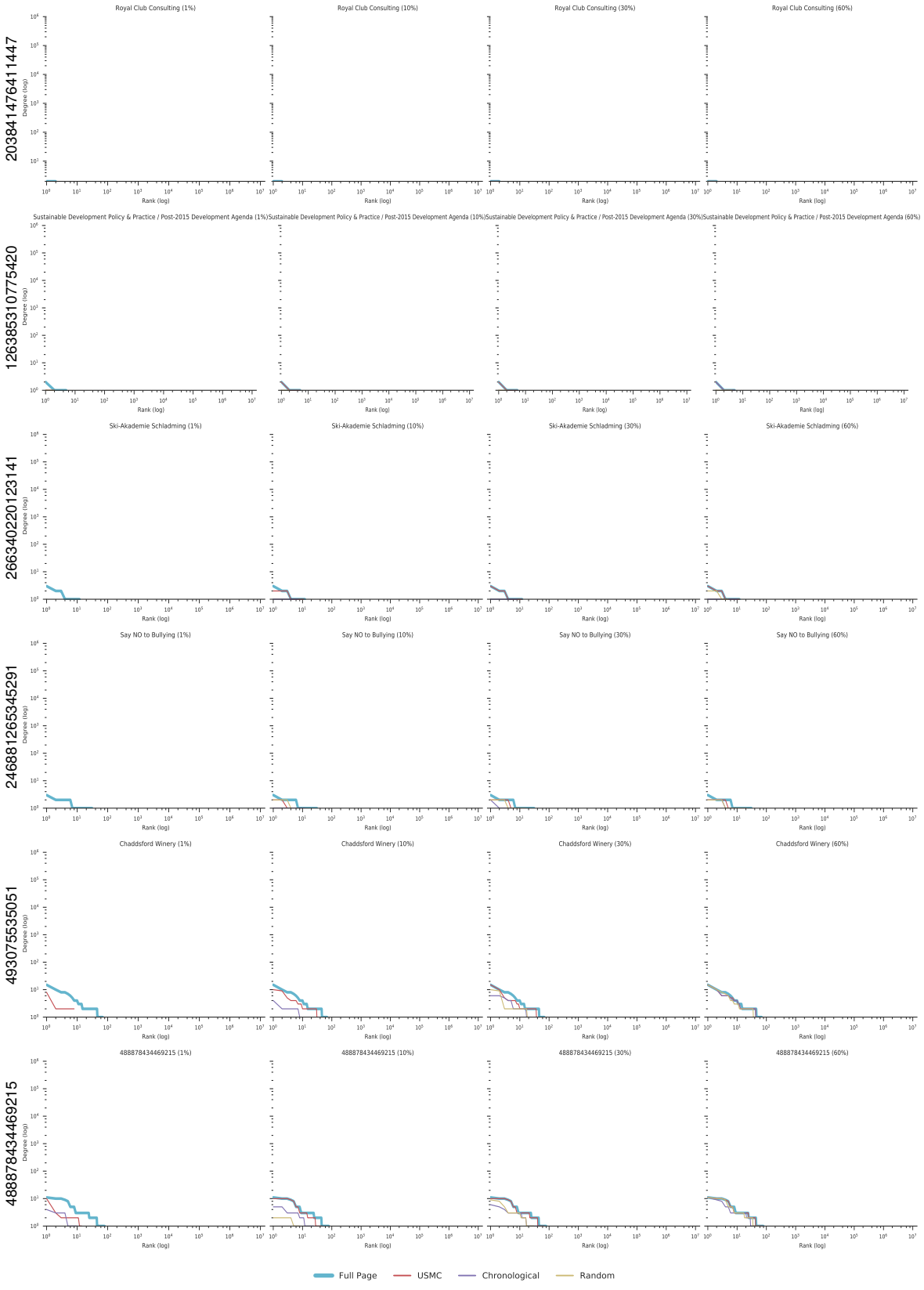


Continued on next page

S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.

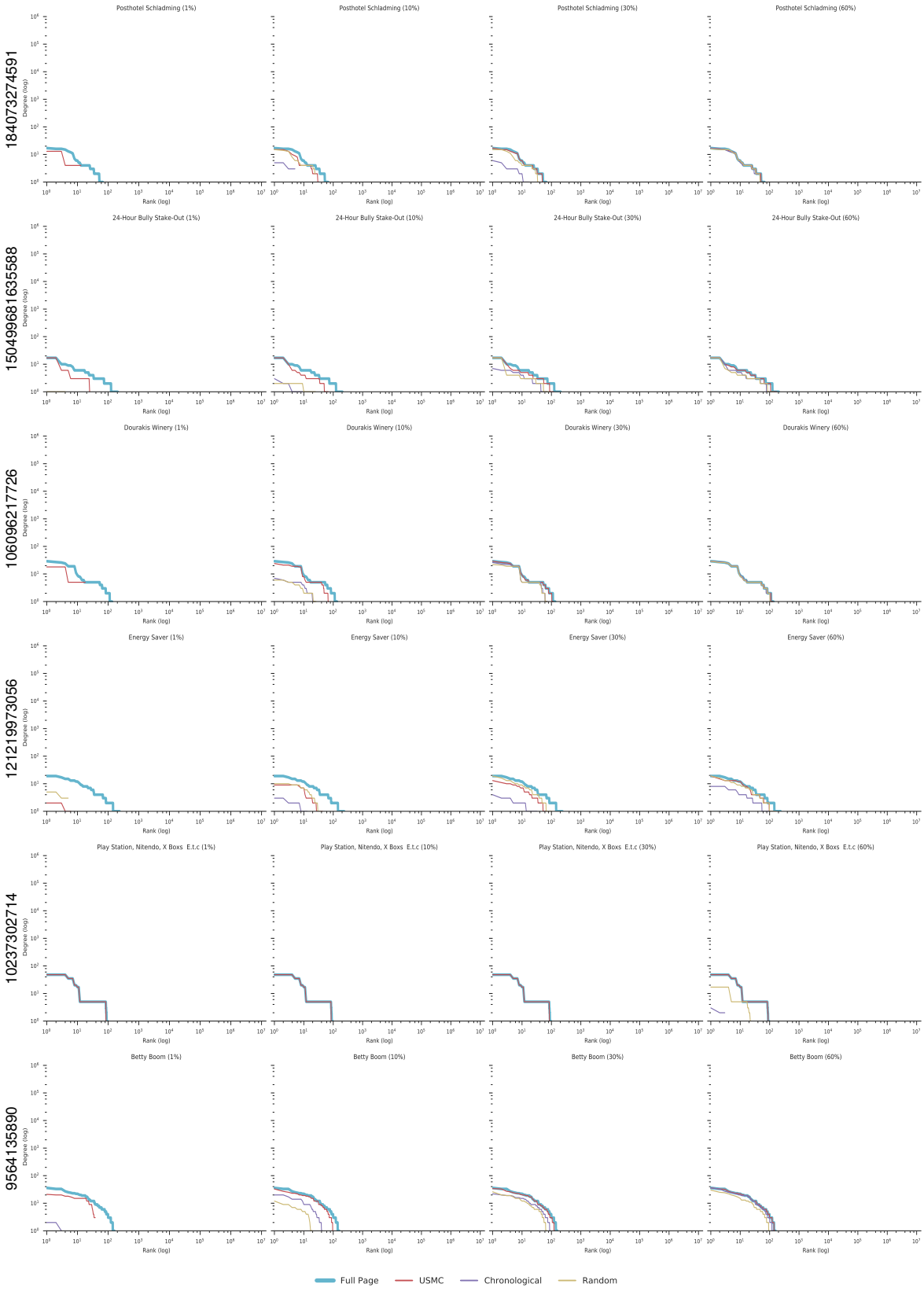


S1 Fig. The effect of ranking with regards to number of interactions of the sample size *left figure* and crawling time *right figure* from a subset of a page's posts ranked by *number of comments*, *number of likes*, *post lifetime* and a *random sample*. The *random sample* shows the median and standard deviation for 100 runs.



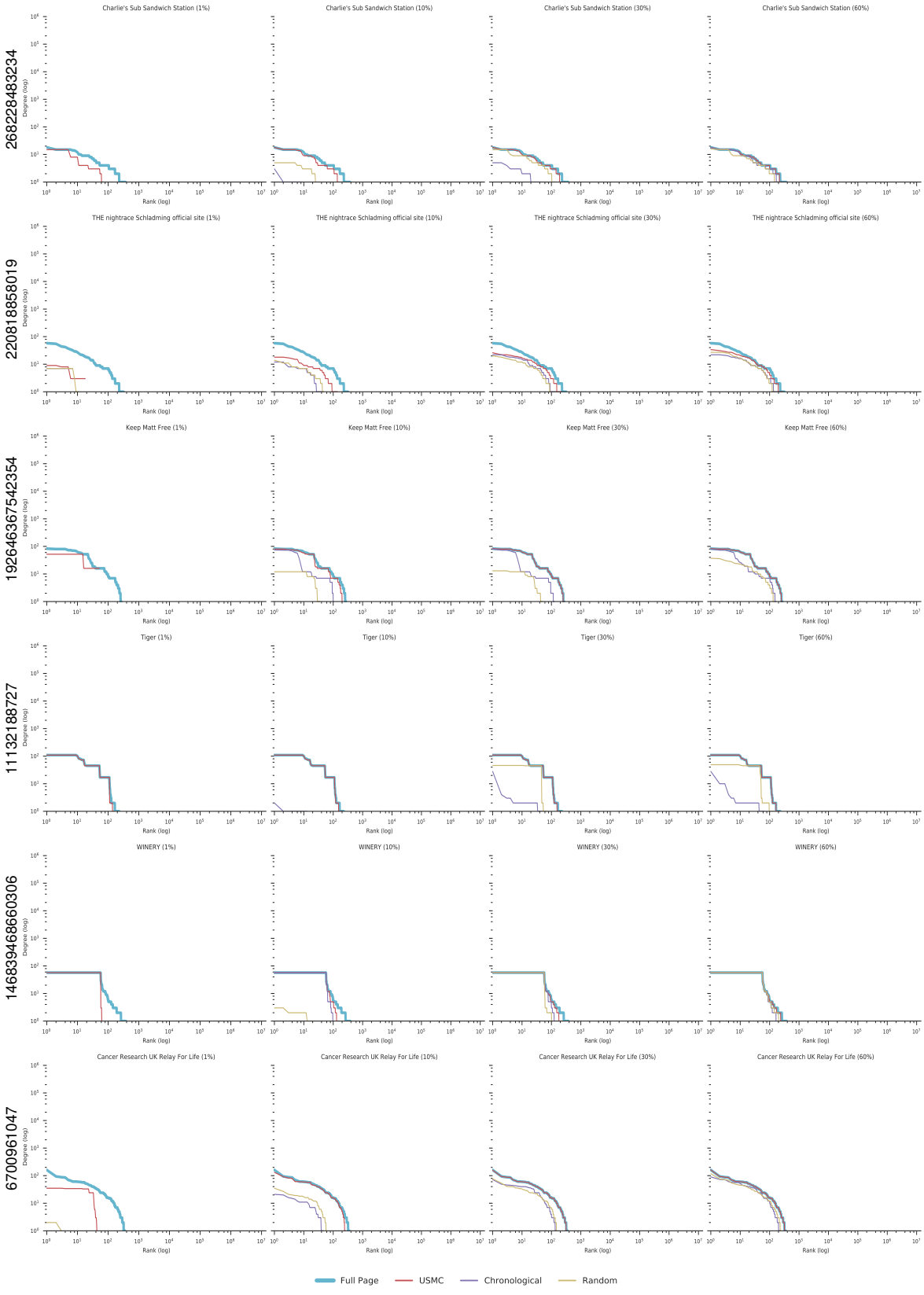
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



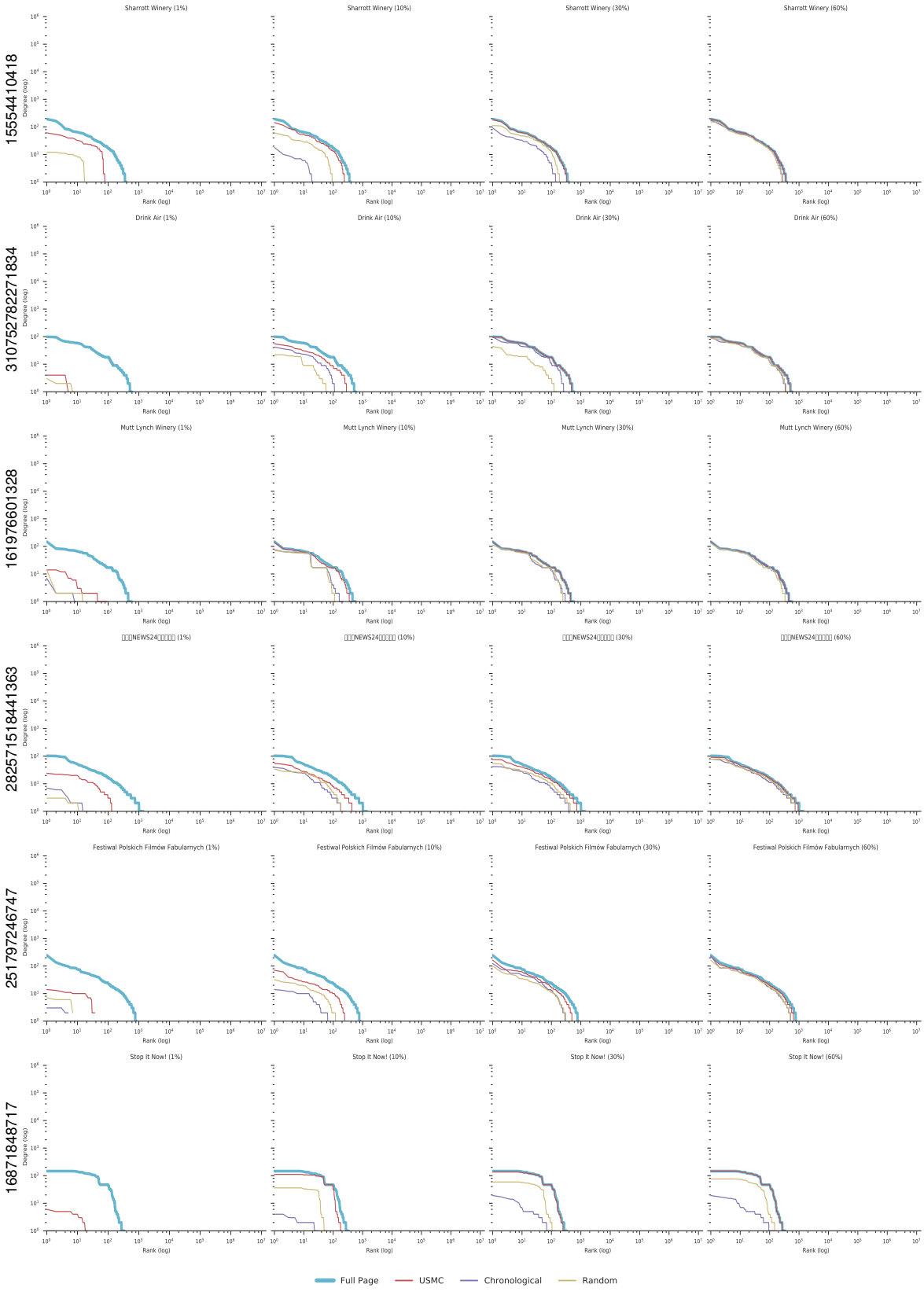
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



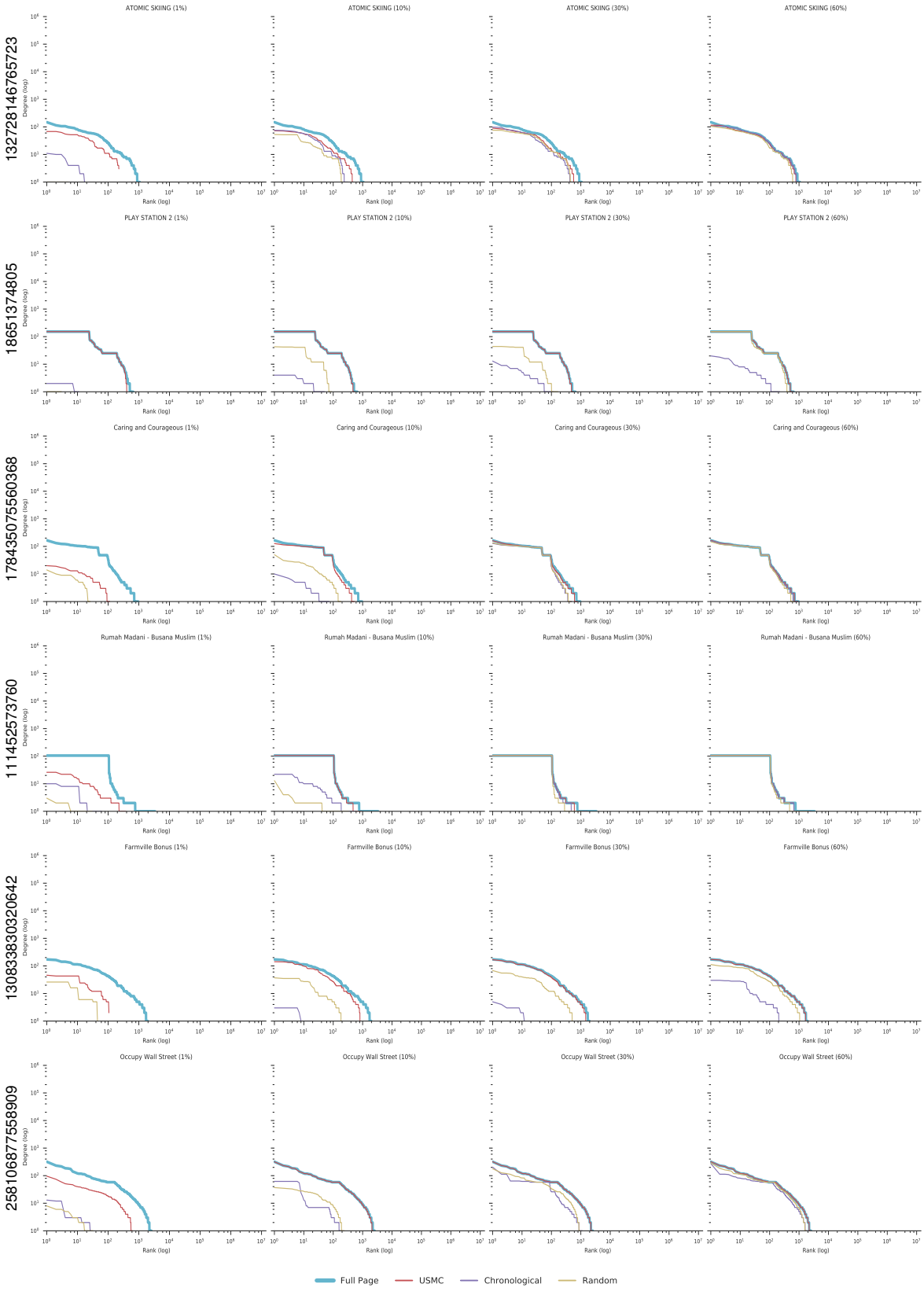
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



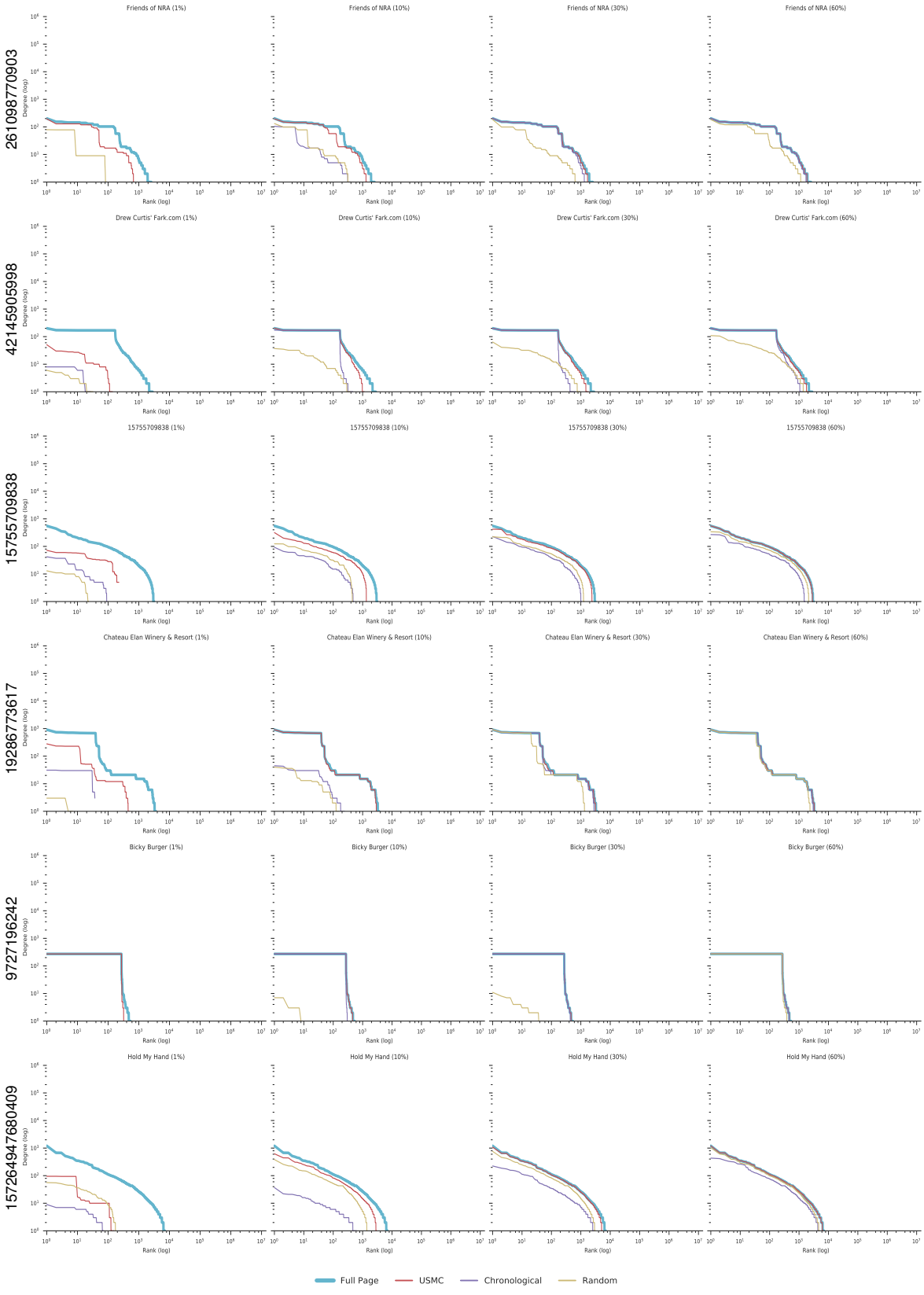
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



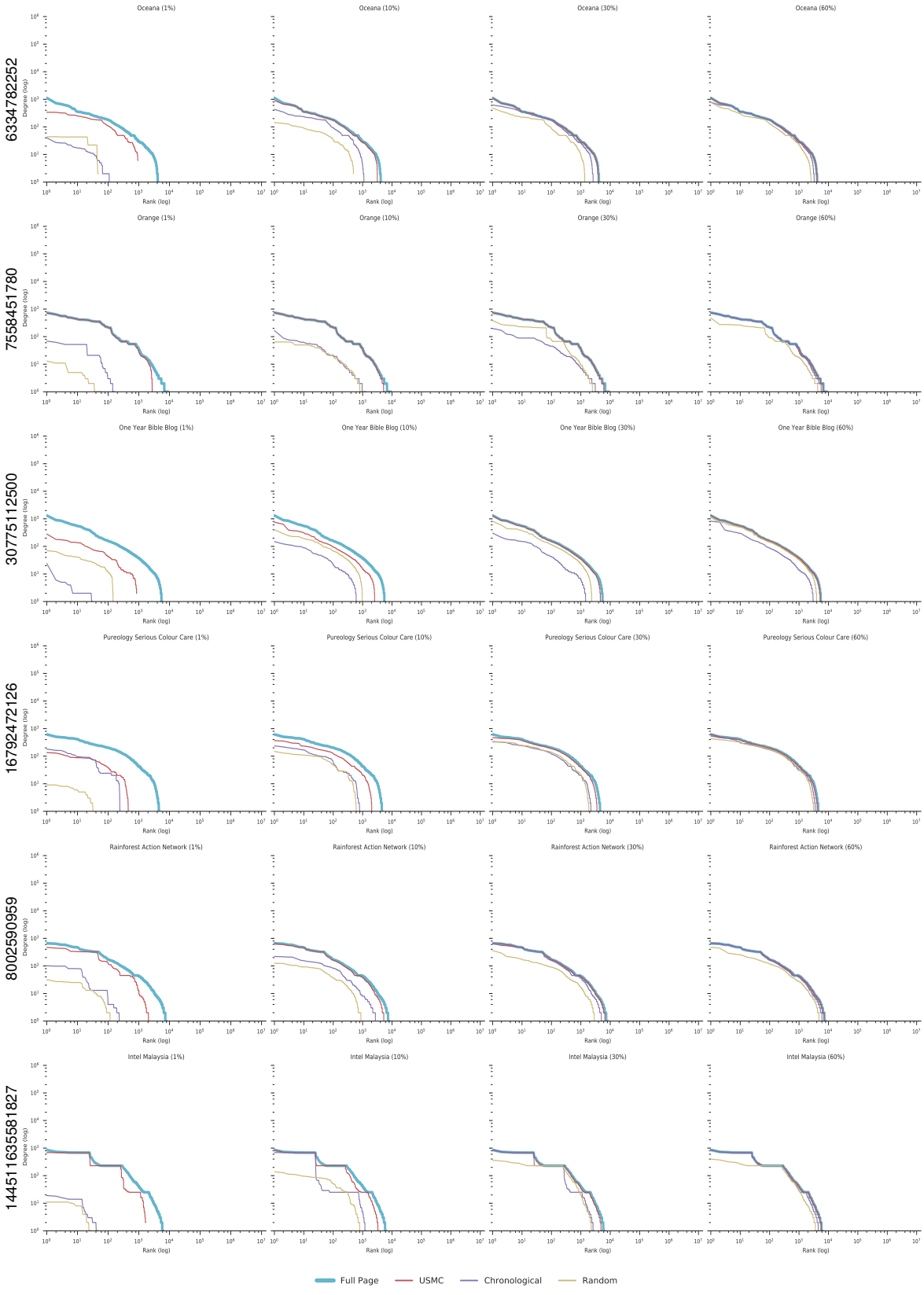
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



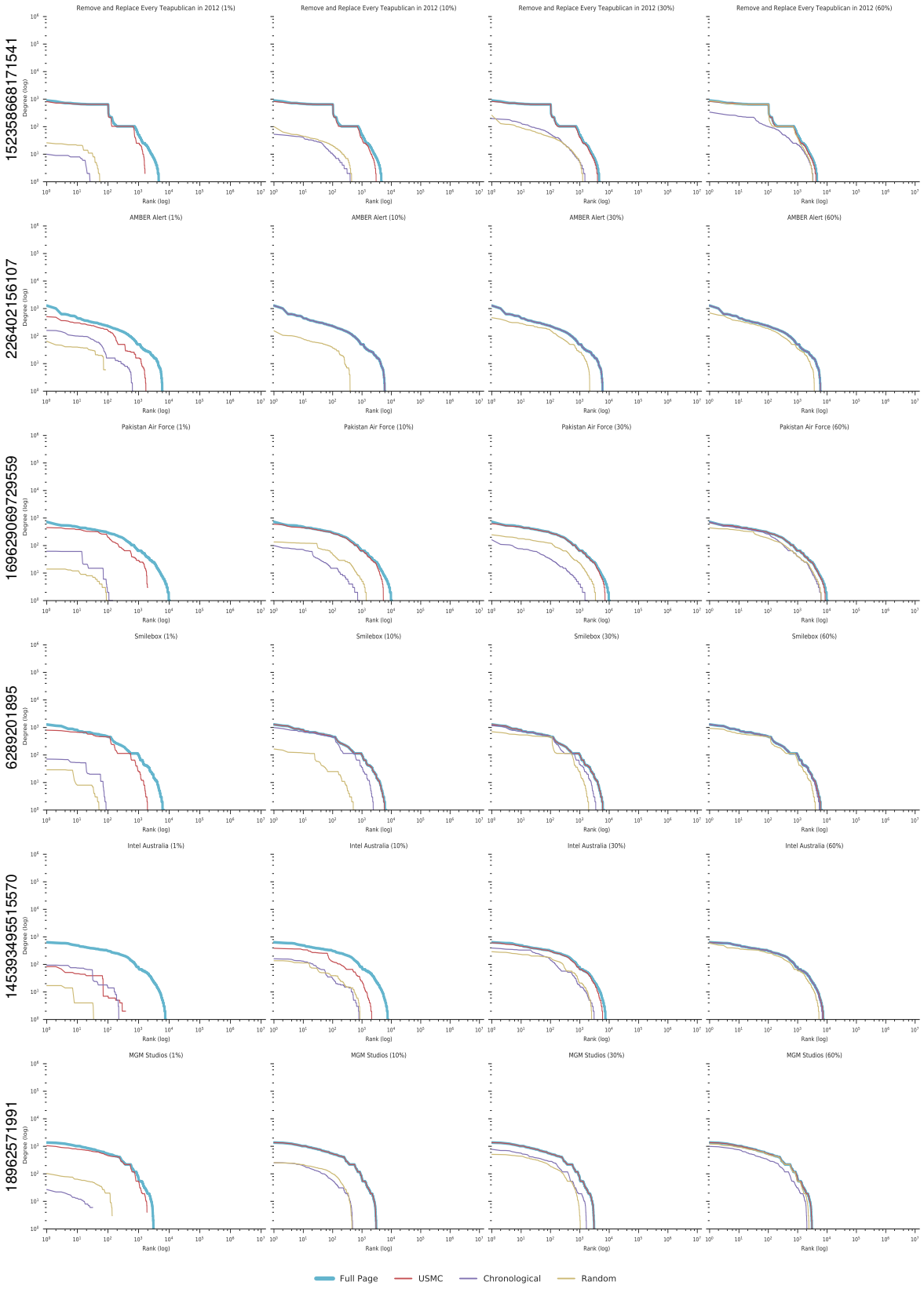
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



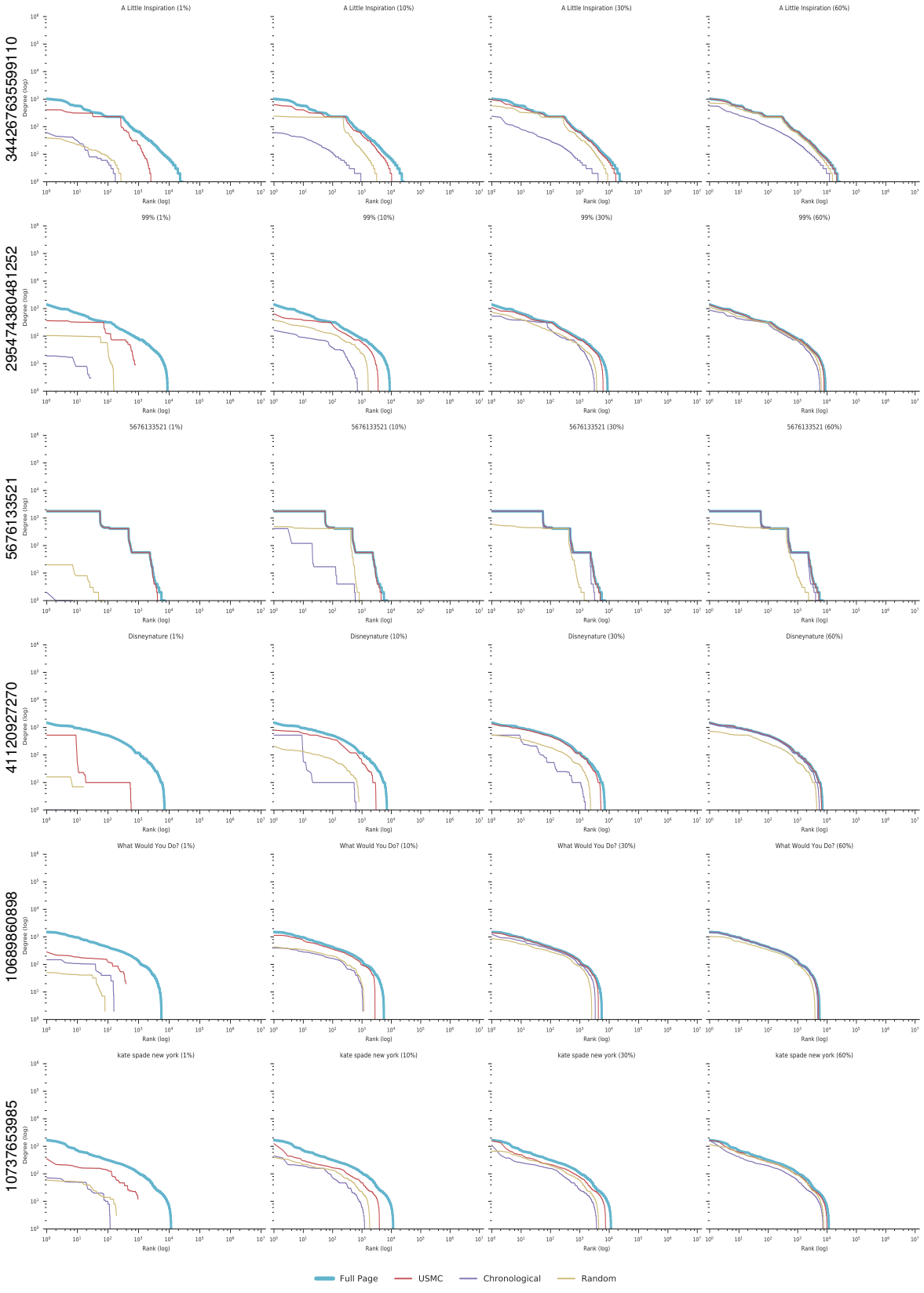
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



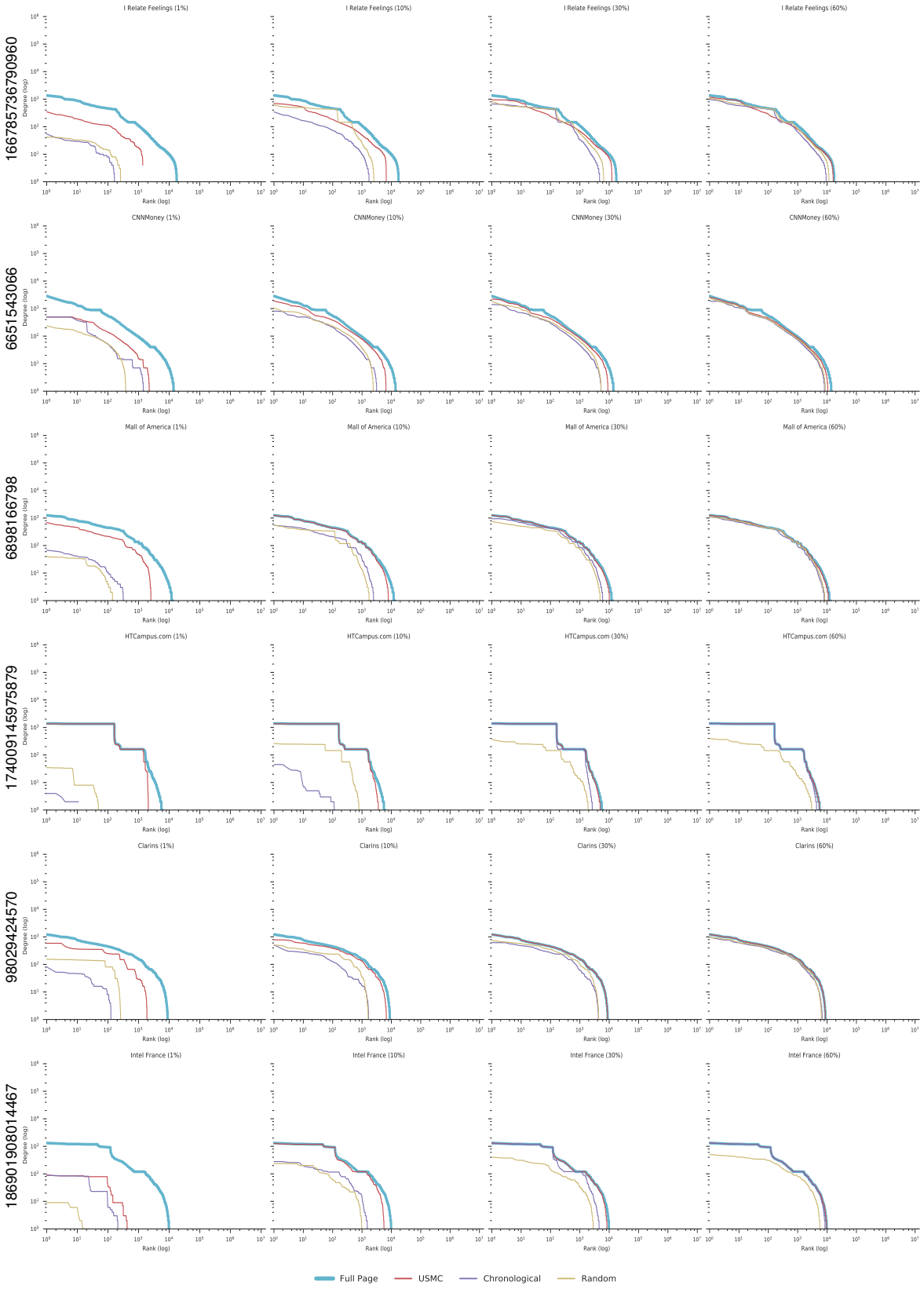
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



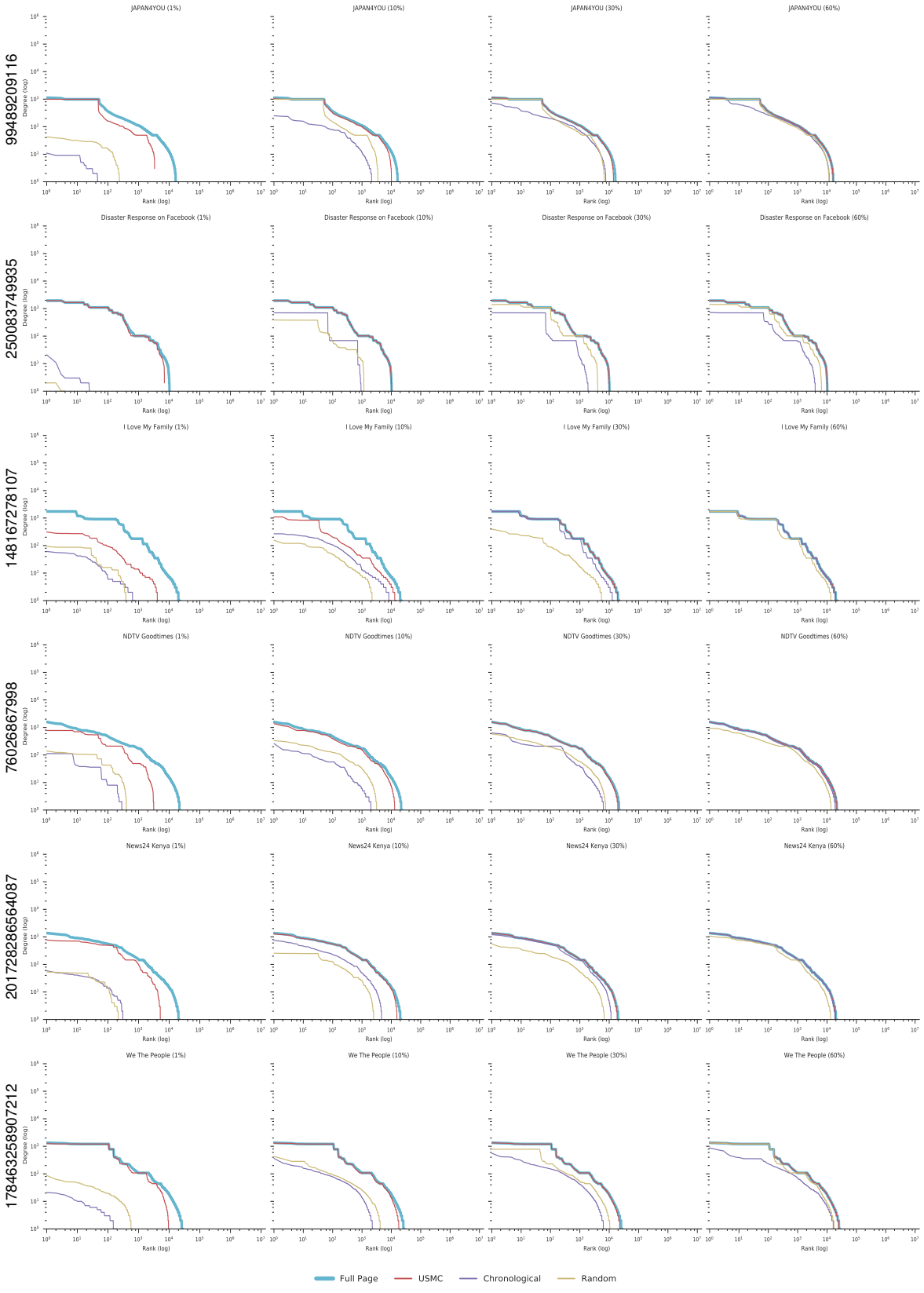
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



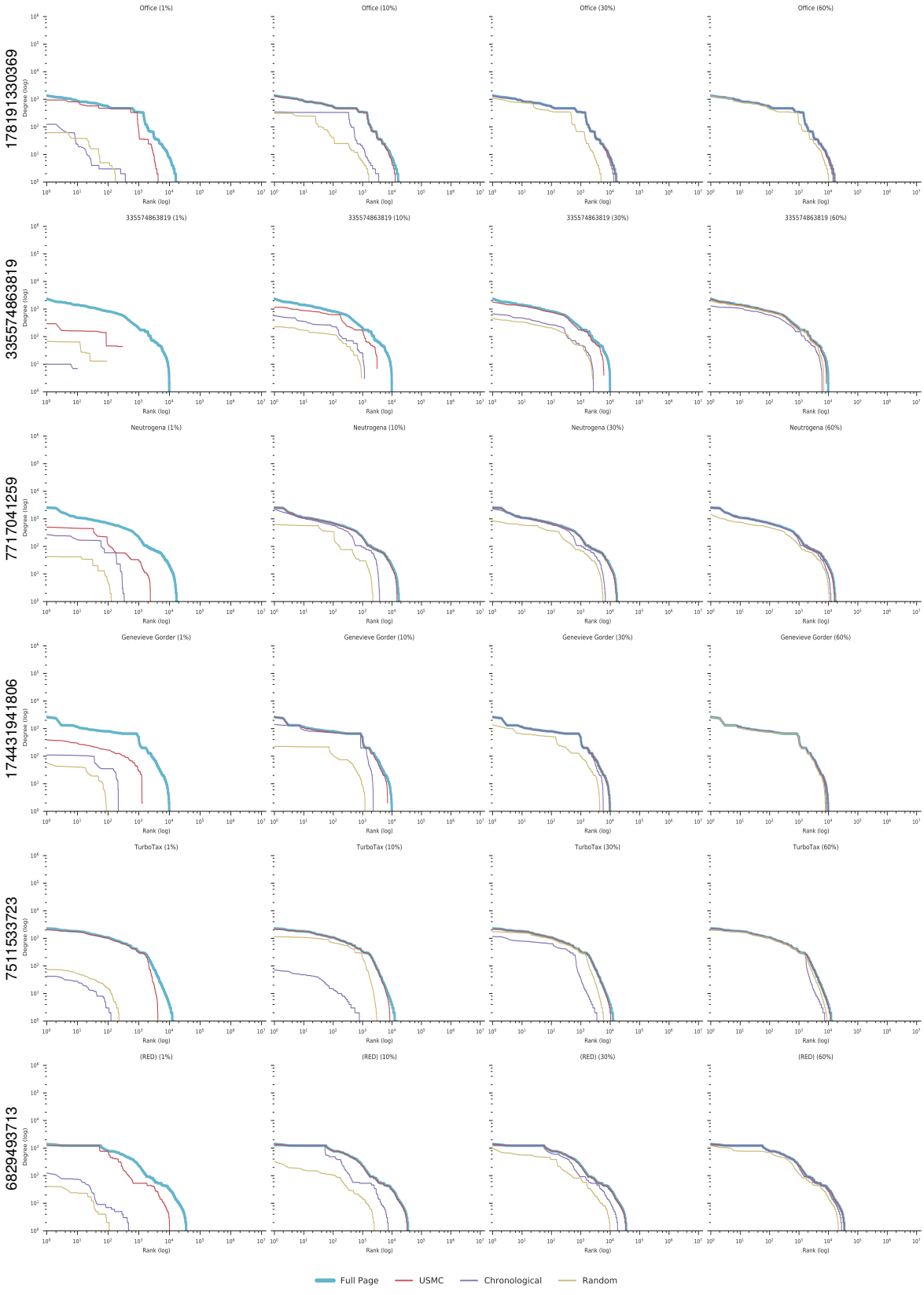
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



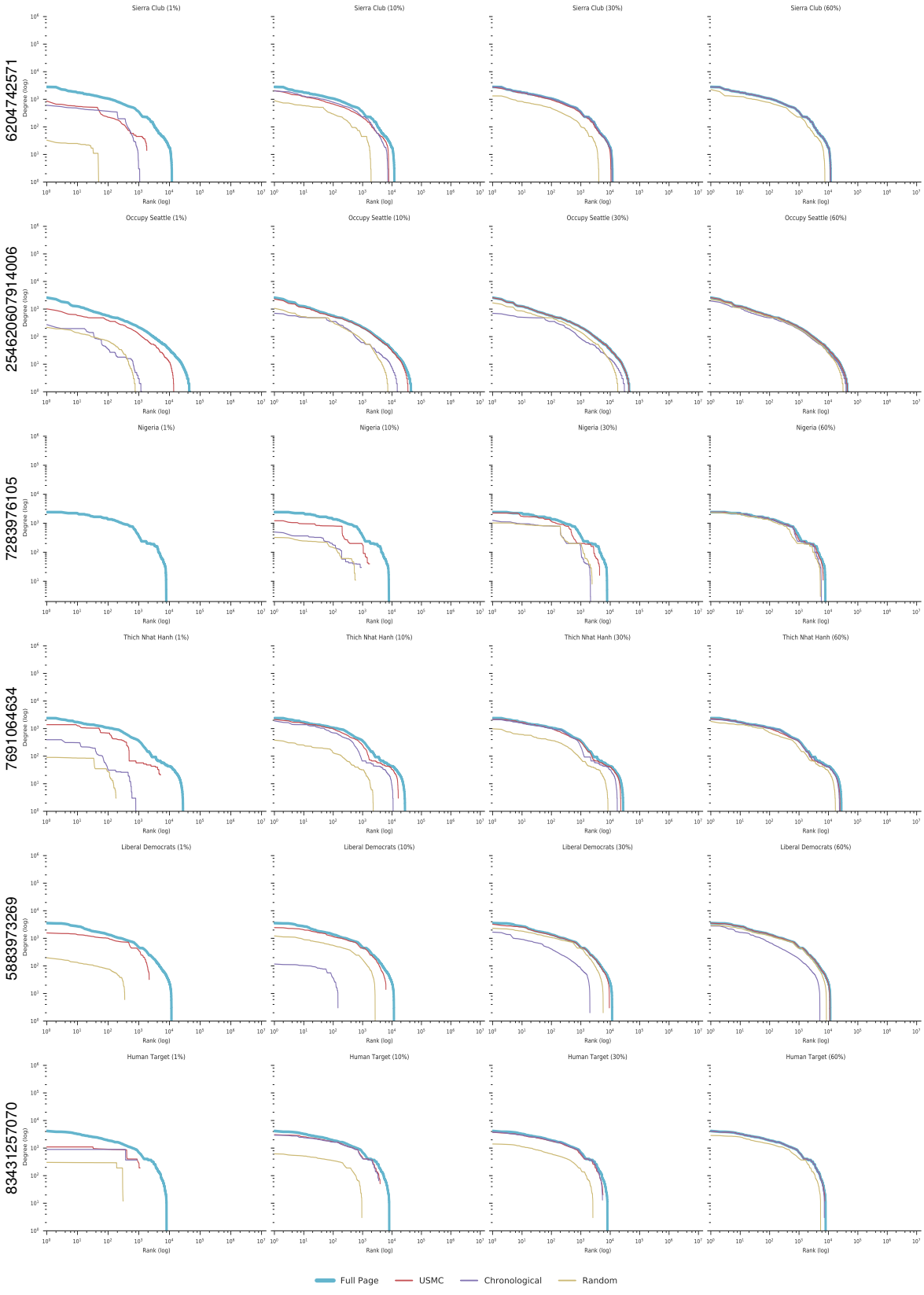
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



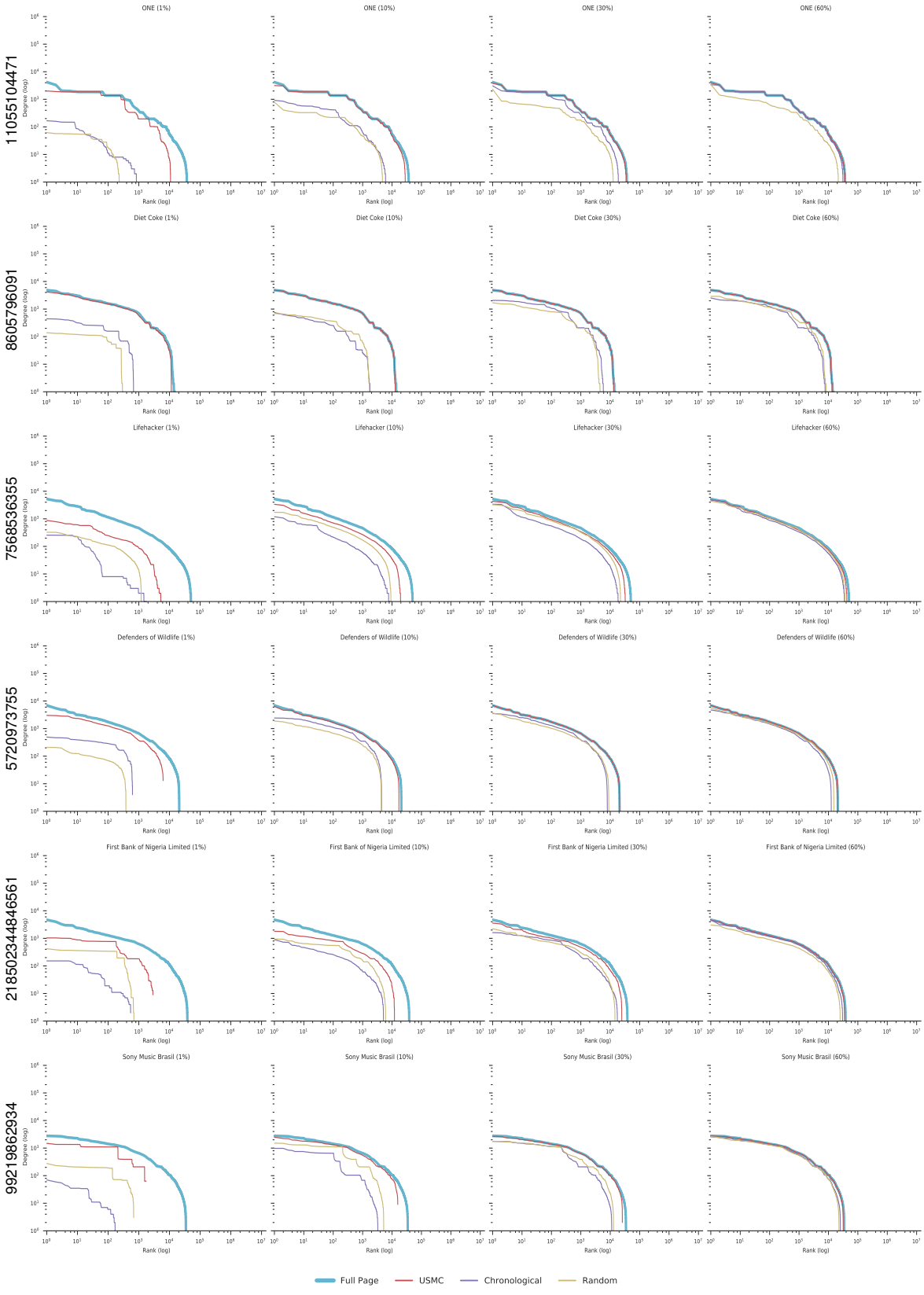
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



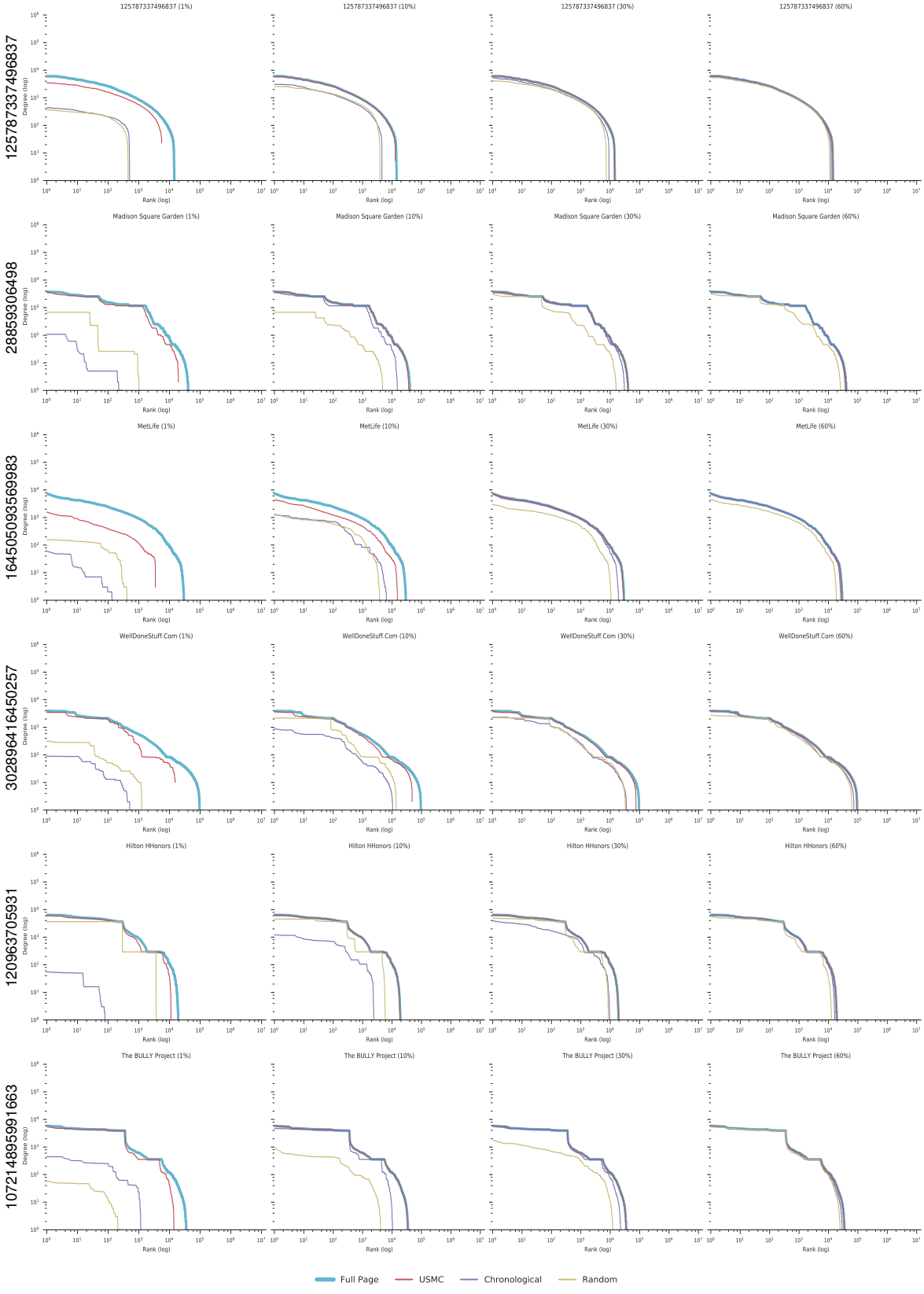
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



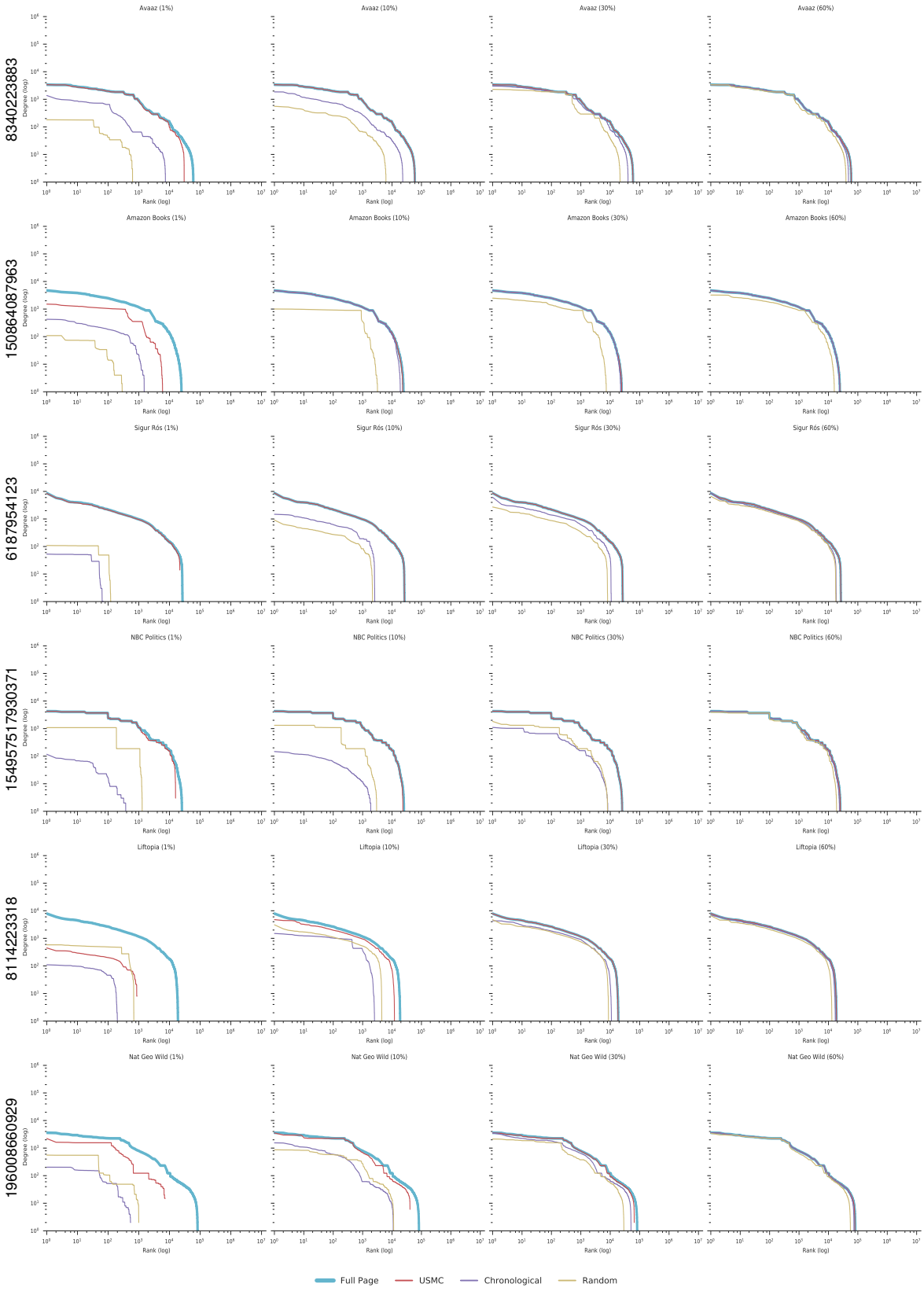
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



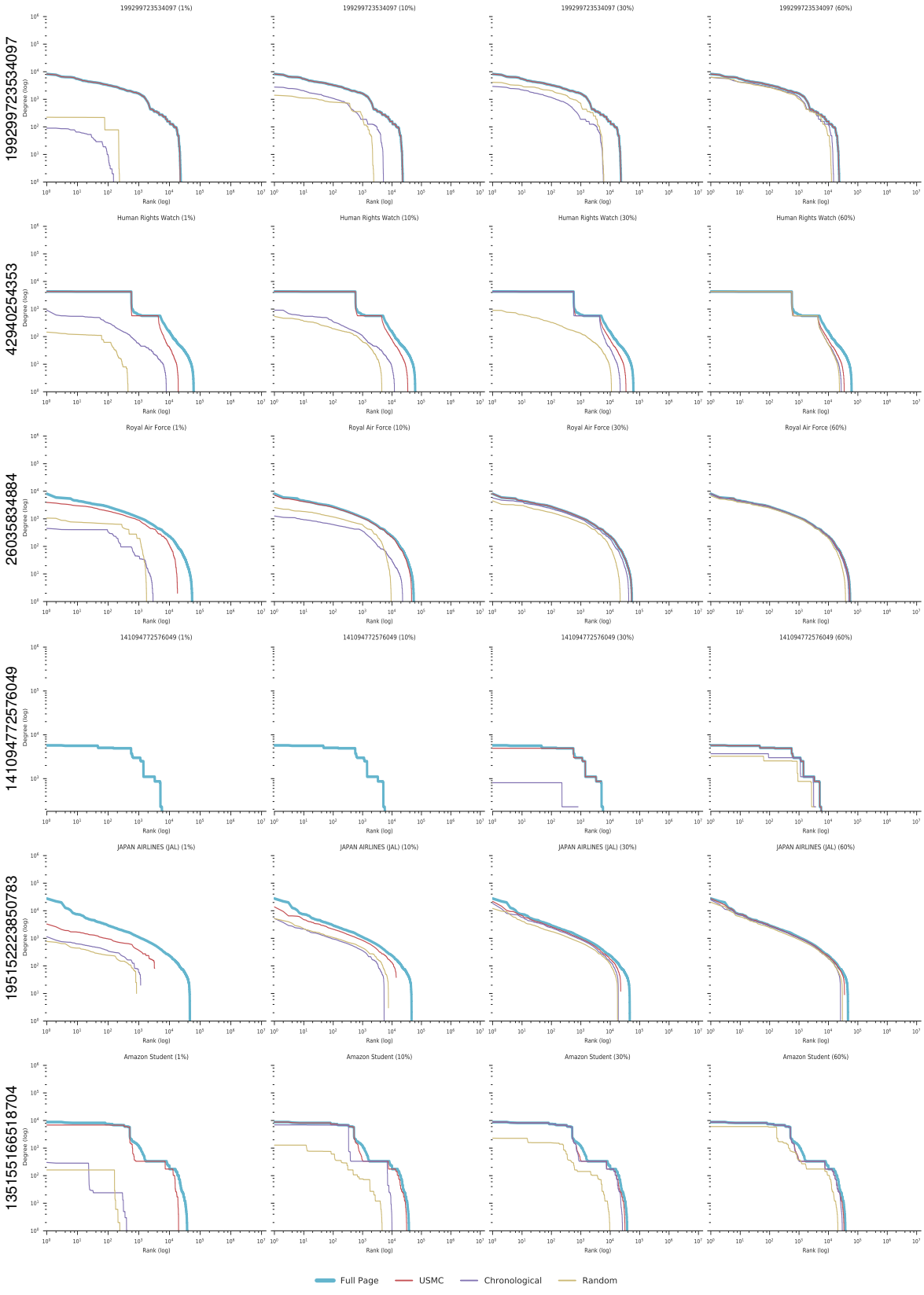
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



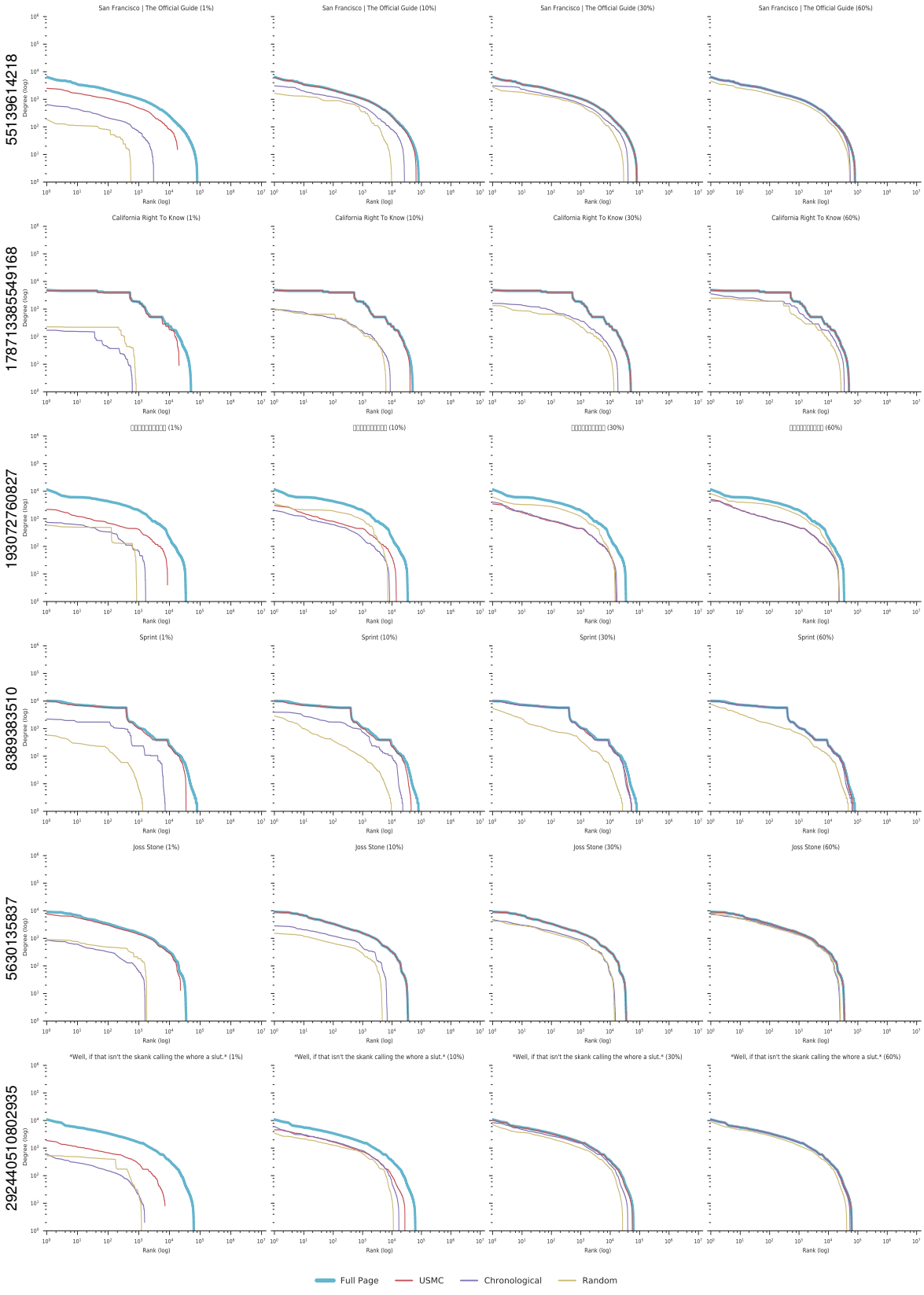
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



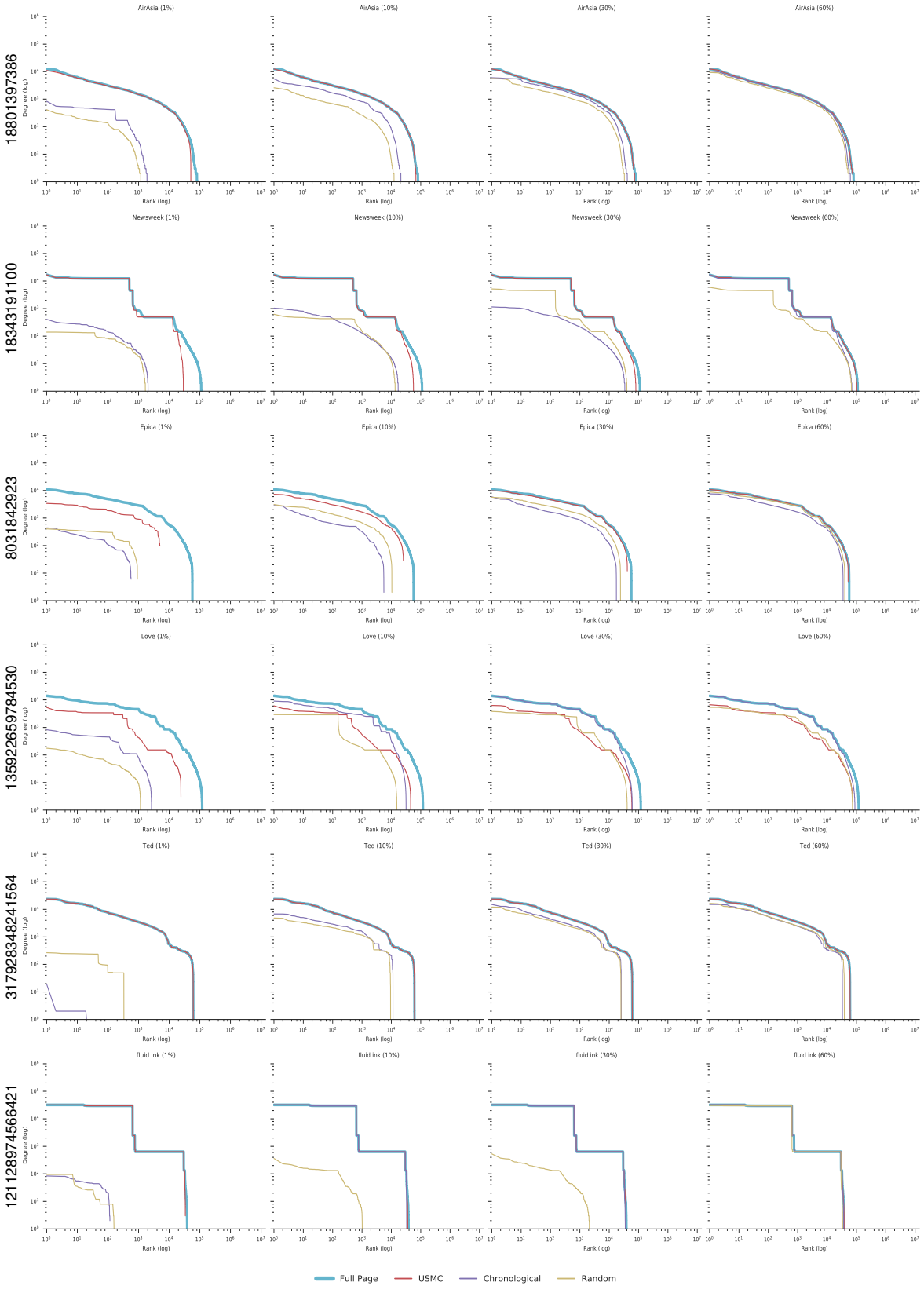
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



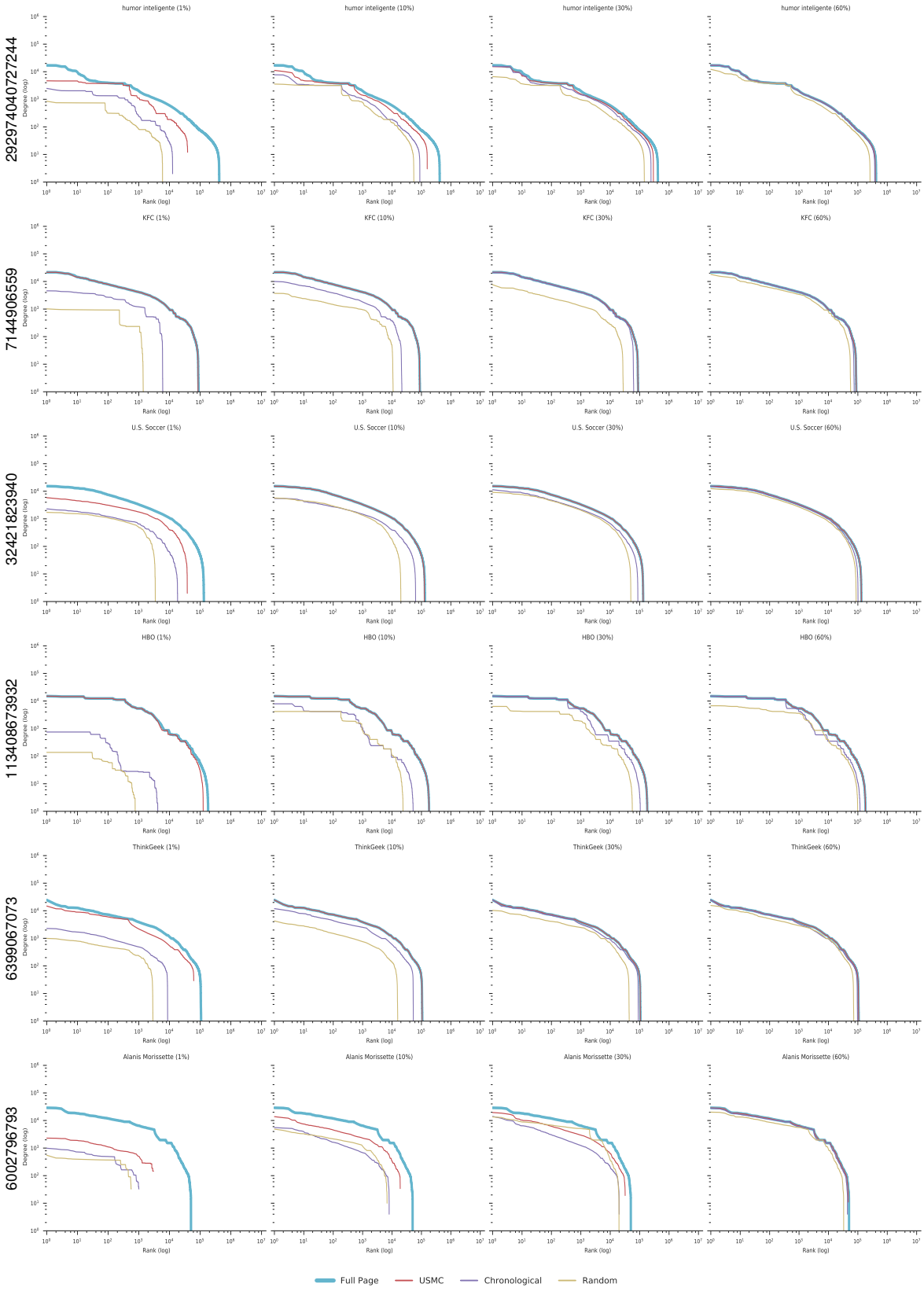
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



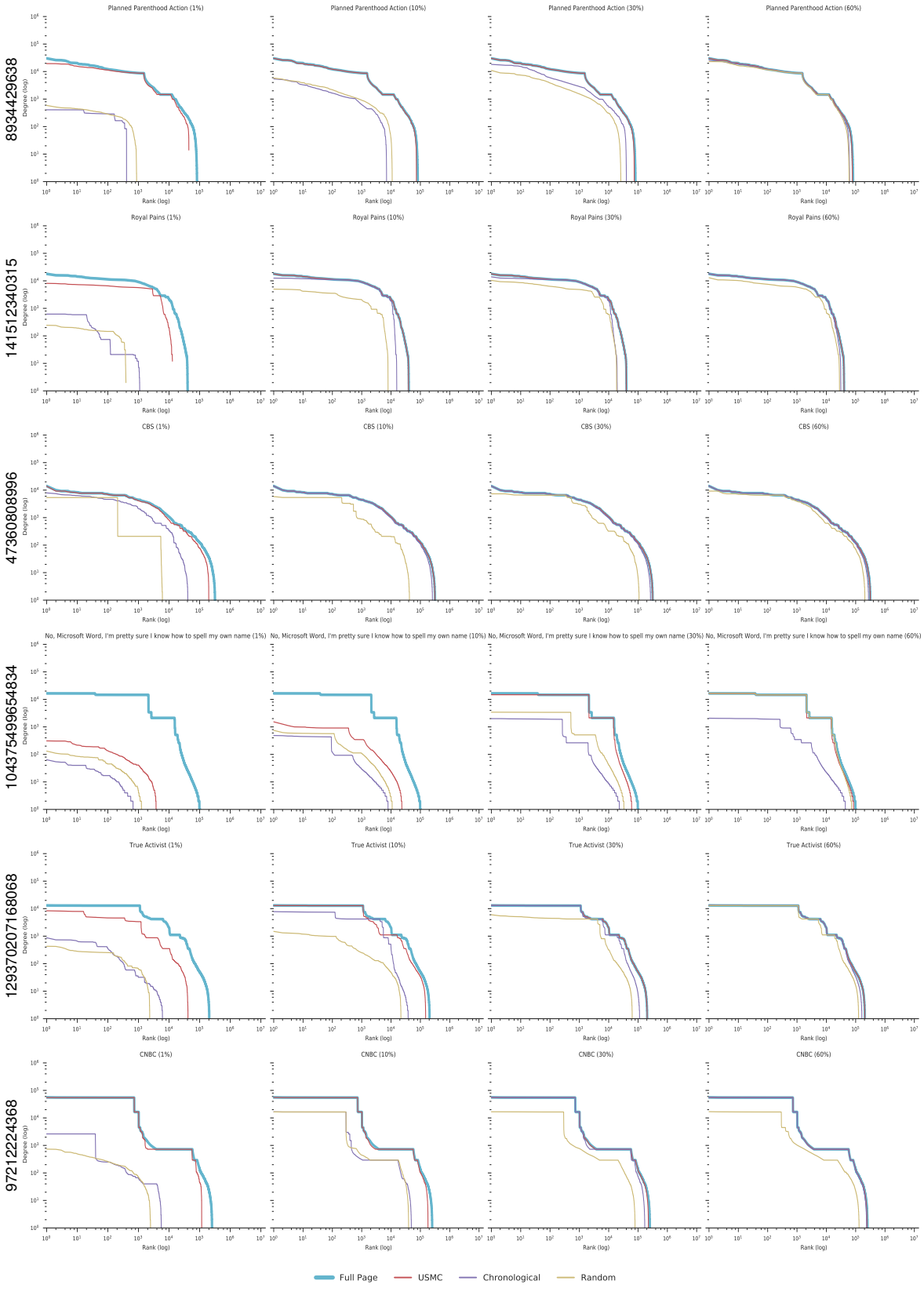
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



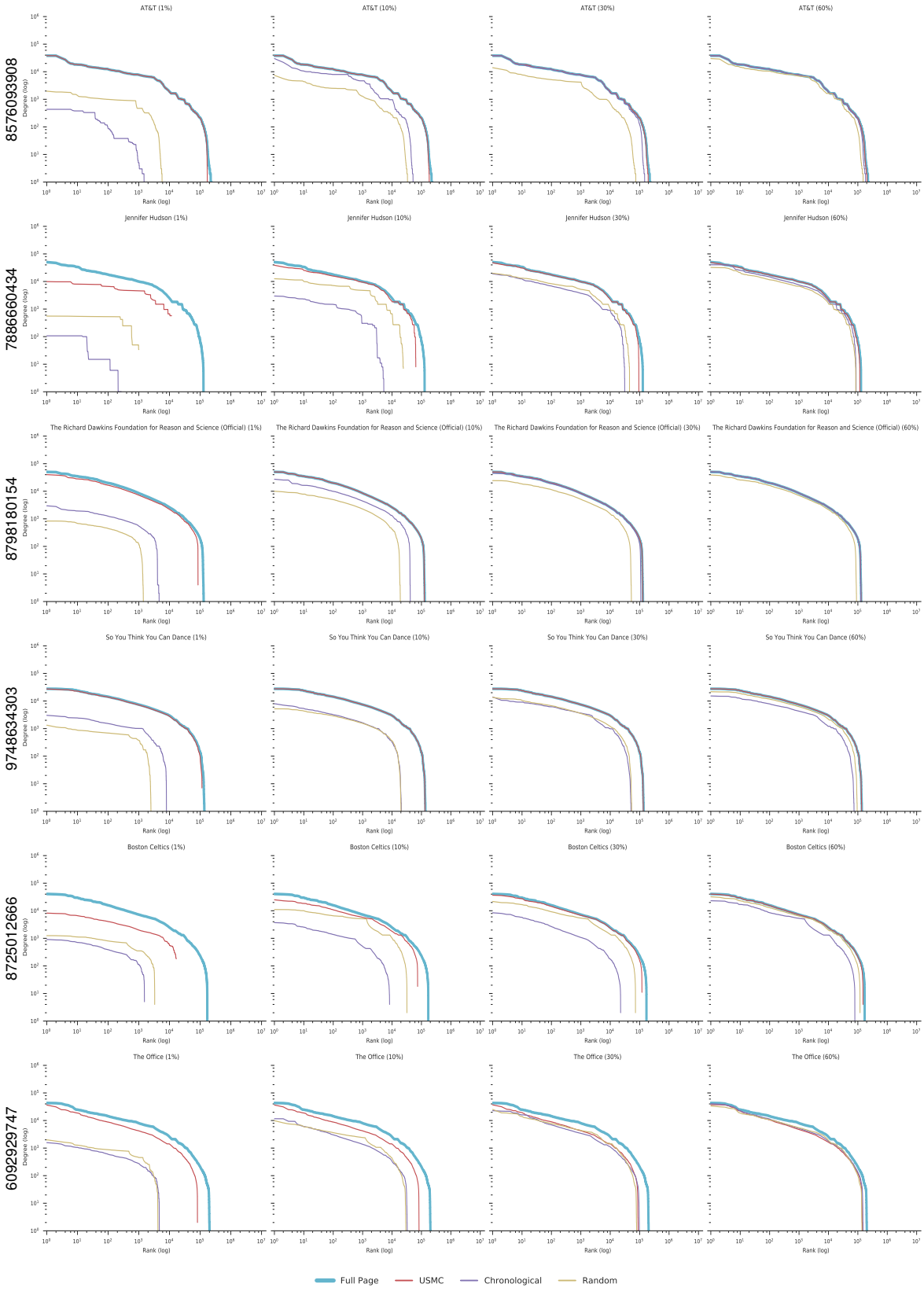
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



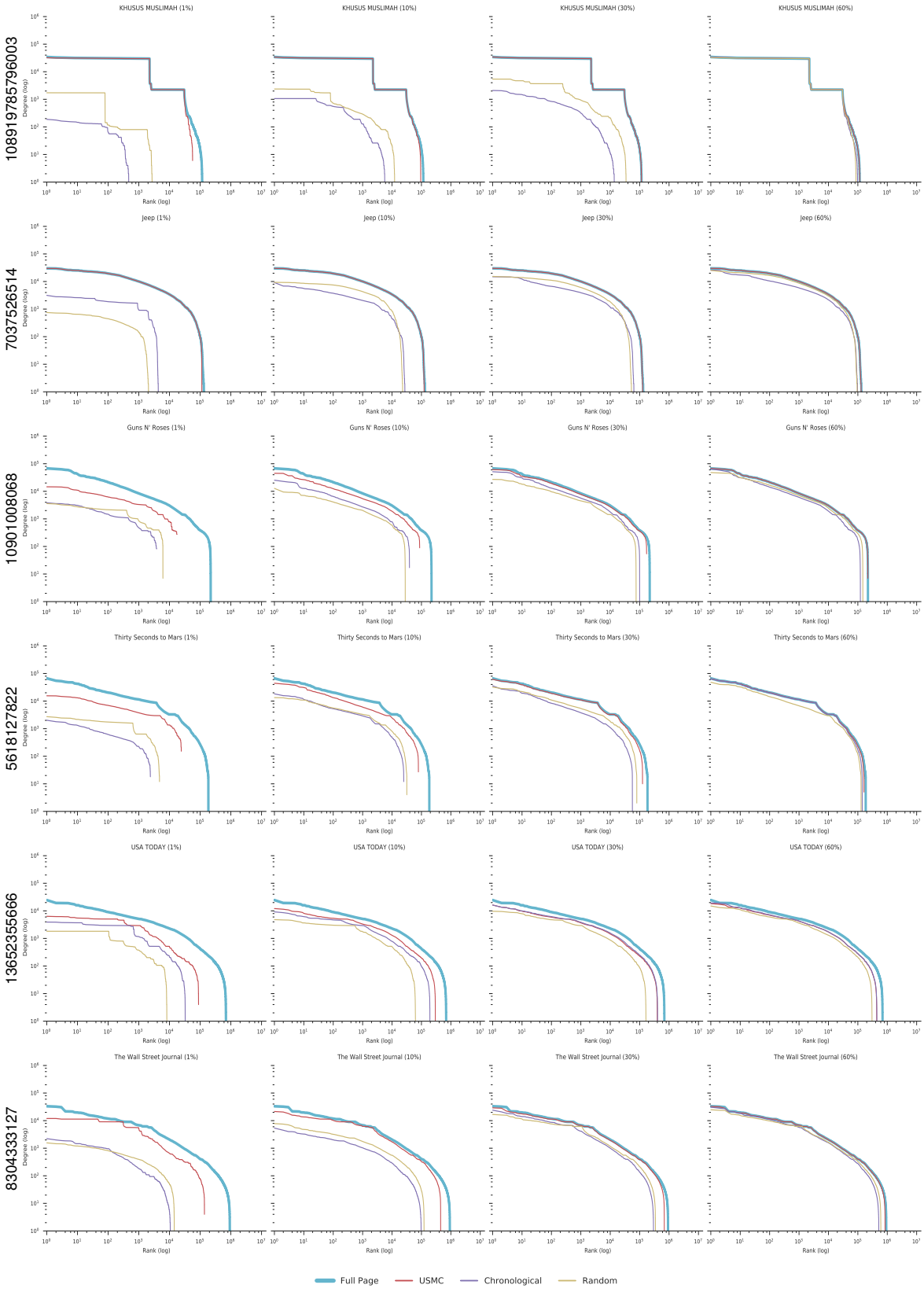
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



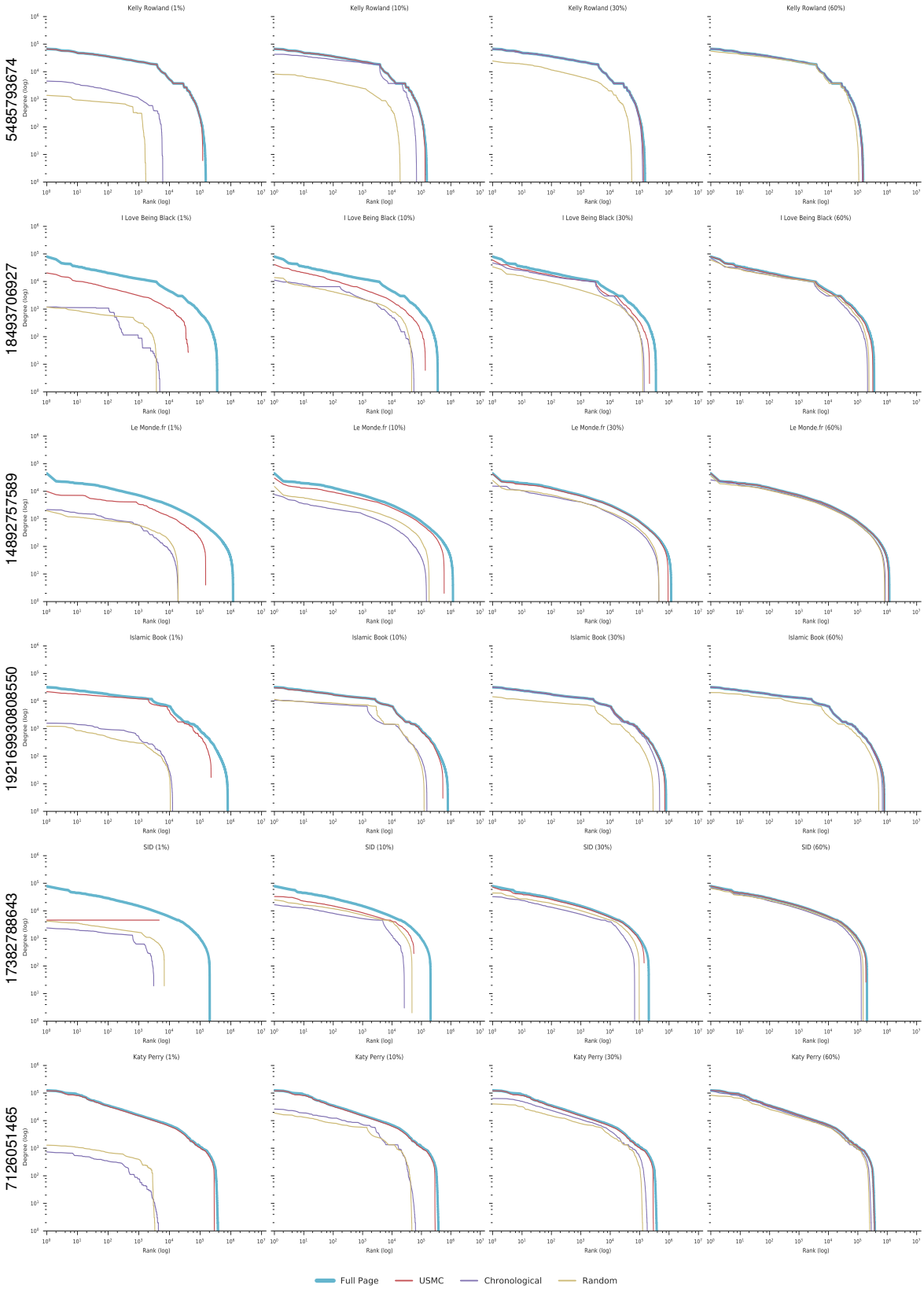
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



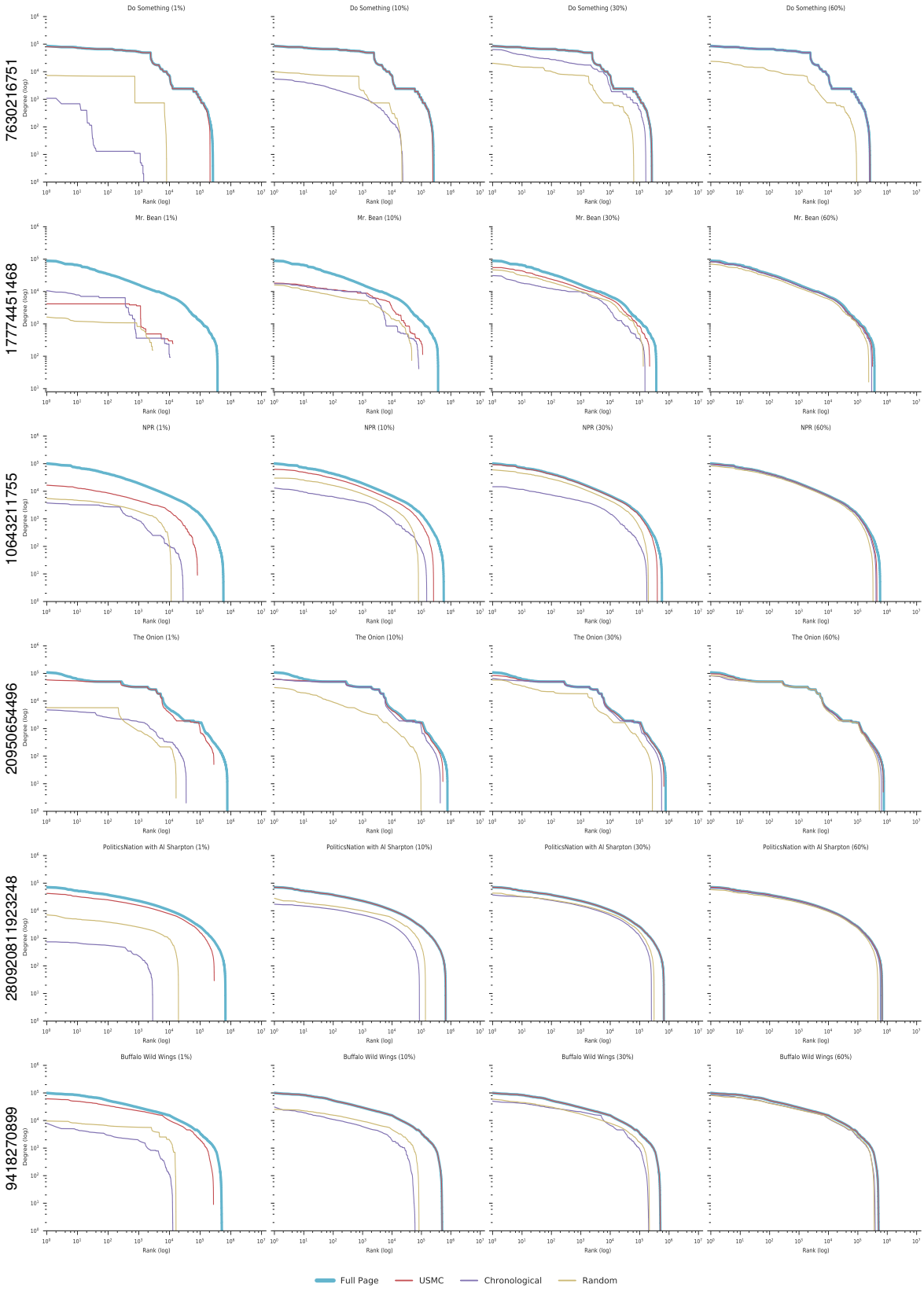
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



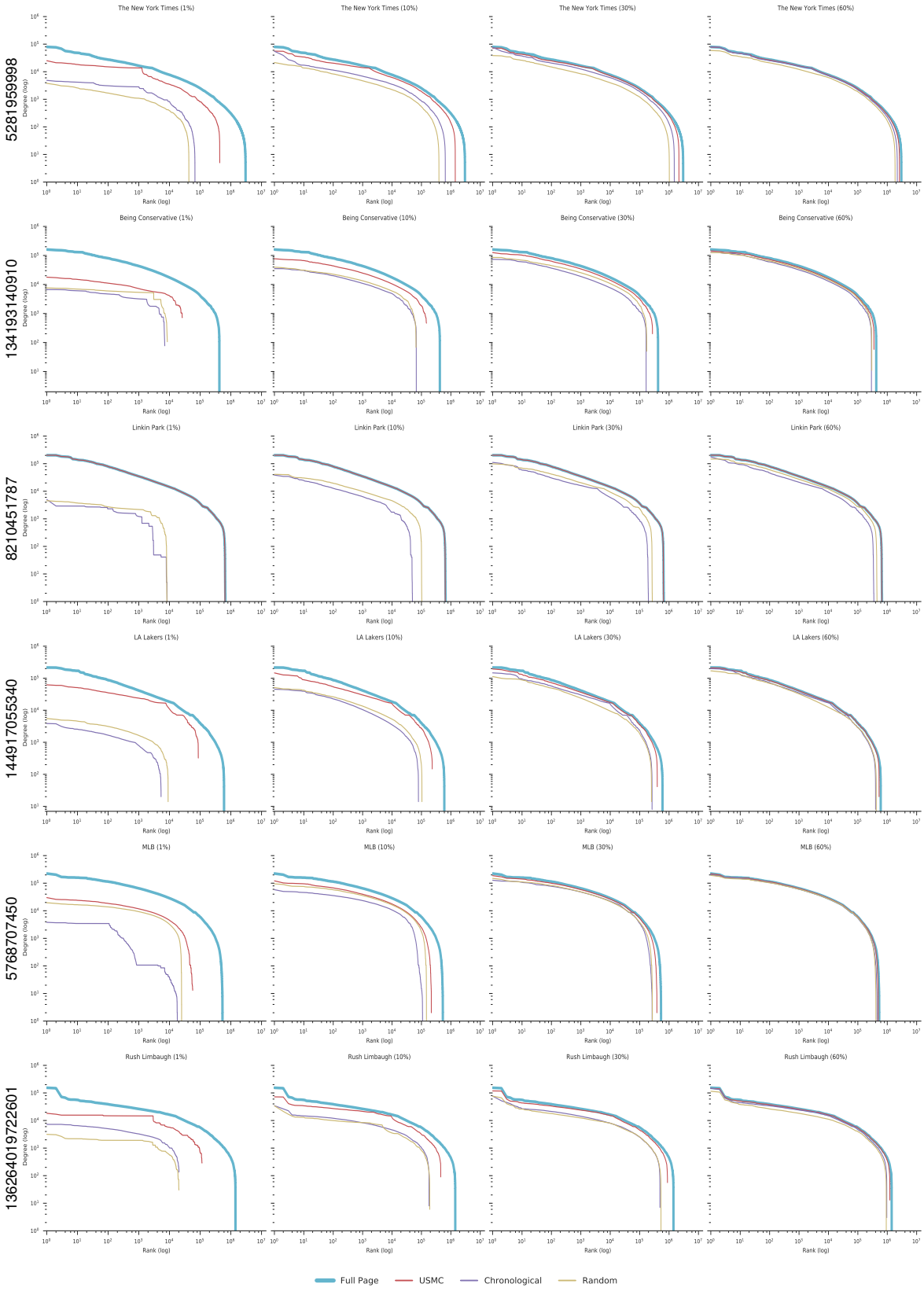
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



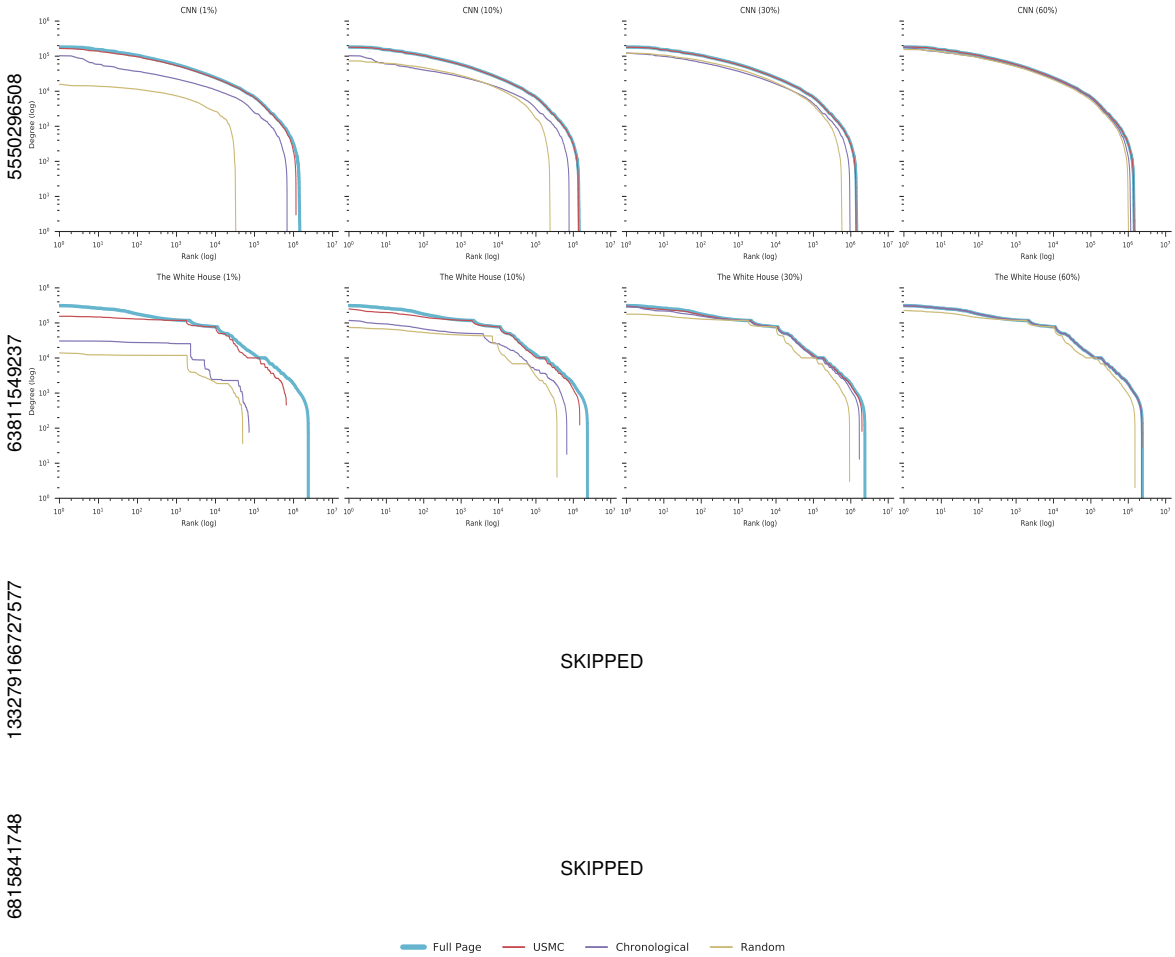
Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



Continued on next page

S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.



S2 Fig. Degree distribution for the analysed Facebook pages. Each column represents a sample size (1%, 10%, 30%, and 60%), and each row represents a page.